CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

Determining the Trends and Most Important World Development Indicators for Life Expectancy at Birth Using World Bank Data

A graduate project submitted in partial fulfillment of the requirements
For the degree of Master of Science in Business Analytics

By

Celine Abrahamian, Joshua Cabal, Namrata Patil, and Christopher Yeung

May 2025

The graduate project of Celine Abrahamian, Joshua Cabal, Namrata Patil, and Christopher Yeung is approved:

Dr. Akash Gupta Date

Dr. Qiuhua Sheng Date

Dr. Pouyan Eslami, Chair Date

California State University, Northridge

Table of Contents

List of Tables

List of Figures

Abstract

Determining the Trends and Most Important World Development Indicators for Life Expectancy at Birth Using World Bank Data

By

Celine Abrahamian, Joshua Cabal, Namrata Patil, and Christopher Yeung

Master of Science in Business Analytics

Life expectancy at birth, as defined by the World Health Organization, represents the average number of years a newborn is expected to live if current age and sex-specific mortality rates persist. The primary objective of this paper is to analyze global trends and identify the most influential development indicators affecting life expectancy at birth. This research examines these trends across various countries, territories, and geographic areas. This study utilizes twenty years of data from the World Bank's World Development Indicators (WDI) Databank (N = 2886, Years: 2004 - 2024, Countries = 171). After an exploratory data analysis to investigate temporal trends and complex patterns in life expectancy, we applied Ordinary Least Squares and K-Means Clustering to different groups of indicators (environmental-related, socioeconomic-related, health-related) to predict life expectancy at birth and form clusters of countries based on commonalities. The health model had the best results with an R-squared of .881, as compared to .806 for the socioeconomic model and .665 for the environmental model. All models identified the coefficients of each indicator to express its strength in affecting life expectancy. Clustering

by indicator group resulted in two or three clusters of common performance of the indicators. The overall results can assist policy makers in understanding where to devote more resources to improve life expectancy outcomes.

Introduction

Life expectancy at birth remains one of the most critical indicators of a nation's overall health and developmental progress ("World Health Statistics"). Defined as the number of years a newborn is expected to live if current mortality trends persist, this metric not only reflects a country's health environment but also encapsulates broader socioeconomic and environmental conditions ("Metadata Glossary"). Over recent decades, an increasing body of research has underscored the importance of socioeconomic factors, environmental quality, and health infrastructure in shaping population health outcomes. Socioeconomic indicators such as income levels, education, and employment opportunities, often determine access to healthcare and influence lifestyle choices, while environmental factors, ranging from air quality to the availability of clean water, directly affect disease prevalence and overall mortality (Office of Disease Preventation and Health Promotion). Health-related indicators, including the incidence of diseases, further compound these outcomes by directly influencing the rate of early mortality and morbidity ("World Health Statistics"). Given the range of contributing factors, a holistic analytical approach is essential for accurately predicting life expectancy trends.

Research Question and Objective

The primary objective of the present research is to build and evaluate traditional prediction models that utilize socioeconomic, environmental, and health-related world development indicators to answer the question of whether it is possible to predict life expectancy at birth using World Bank data. This study aims to identify the most significant predictors and quantify their respective impacts on life expectancy at birth. The research also aims to identify the trends over time for these significant predictors. In doing so, the analysis not only contributes

to academic research on health outcomes but also provides practical insights for decision makers who have a stake in public health strategies.

Methodology

The code we developed and utilized for this study is available in our group's public GitHub repository (Group 1). The study begins by combining data from multiple sources and integrating the data into a normalized database of country-level indicators. After merging and cleaning (imputing missing values with KNN, winsorizing extreme outliers, and standardizing variables), indicators were split into one of three categories: environmental, socioeconomic, or health. For each indicator set, a stepwise feature selection using VIF and p-values was used to refine the final indicators before fitting Ordinary Least Squares (OLS) models. Finally, k-means clustering at five-year intervals from 2005 onward tracked how these cluster centroids and their associated life expectancy patterns evolved over time. This approach aims to offer a simple and comprehensive view of key life expectancy drivers and how their effects shift in diverse global contexts. Figures 1 and 2 summarize these processes on the next pages.

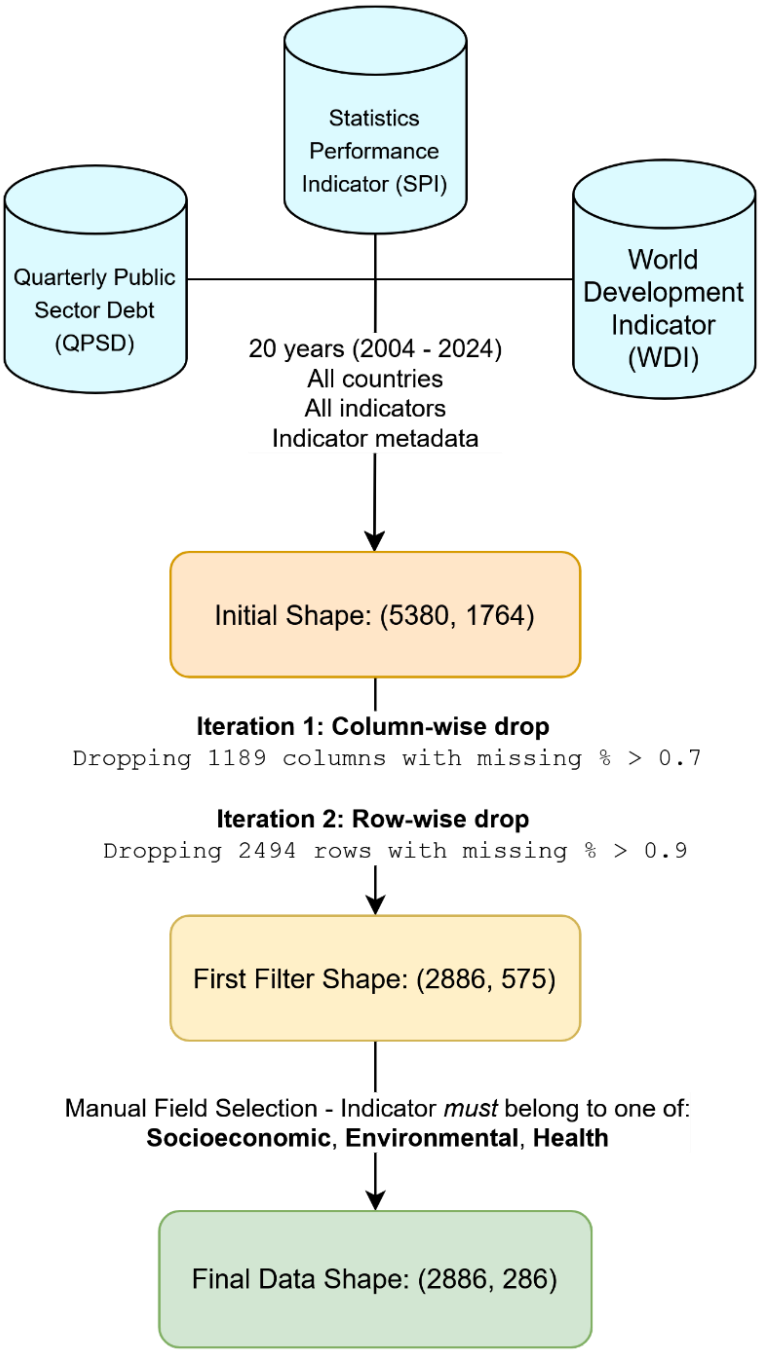


Figure 1. Data Pipeline

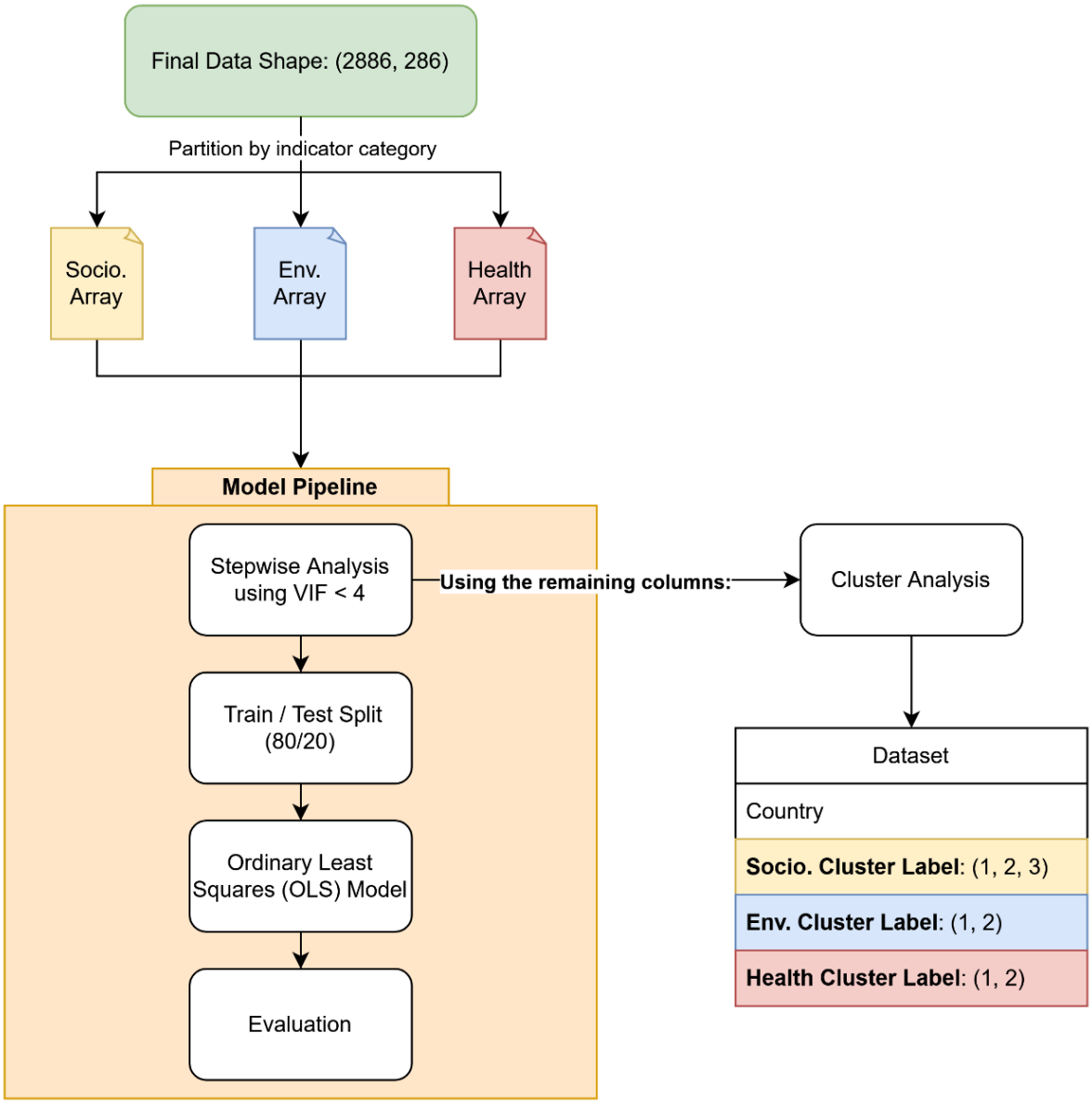


Figure 2. Analysis Pipeline

Data Source

Initially, the data from three databases were extracted: World Development Indicators (WDI), Quarterly Public Sector Debt (QPSD), and Statistical Performance Indicators (SPI). The World Development Indicators is the primary World Bank collection of development indicators,

compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates ("World Development Indicators"). Statistical Capacity Indicators provides information on various aspects of national statistical systems of developing countries, including an overall country-level statistical capacity indicator ("Statistical Performance Indicators"). The Quarterly Public Sector Debt was jointly developed by the World Bank and the International Monetary Fund, and the database brings together detailed public sector debt data of selected developing / emerging market countries ("Quarterly Public Sector Debt").

The data values for all indicators from all countries spanning over the period 2004 - 2024 were exported from these databases. In addition, the metadata of the countries and the indicators were extracted and stored in the database. The exported data files were saved locally then were used together to assemble a normalized database. The country, indicator, and indicator record values for each country-year were stored into one of three respective tables: Country (1 table), DB.TopicIndicator (18 total tables), DB.Record (18 total tables). The structure of the database tables and the configured relationships are shown in Figure 3 on the next page.

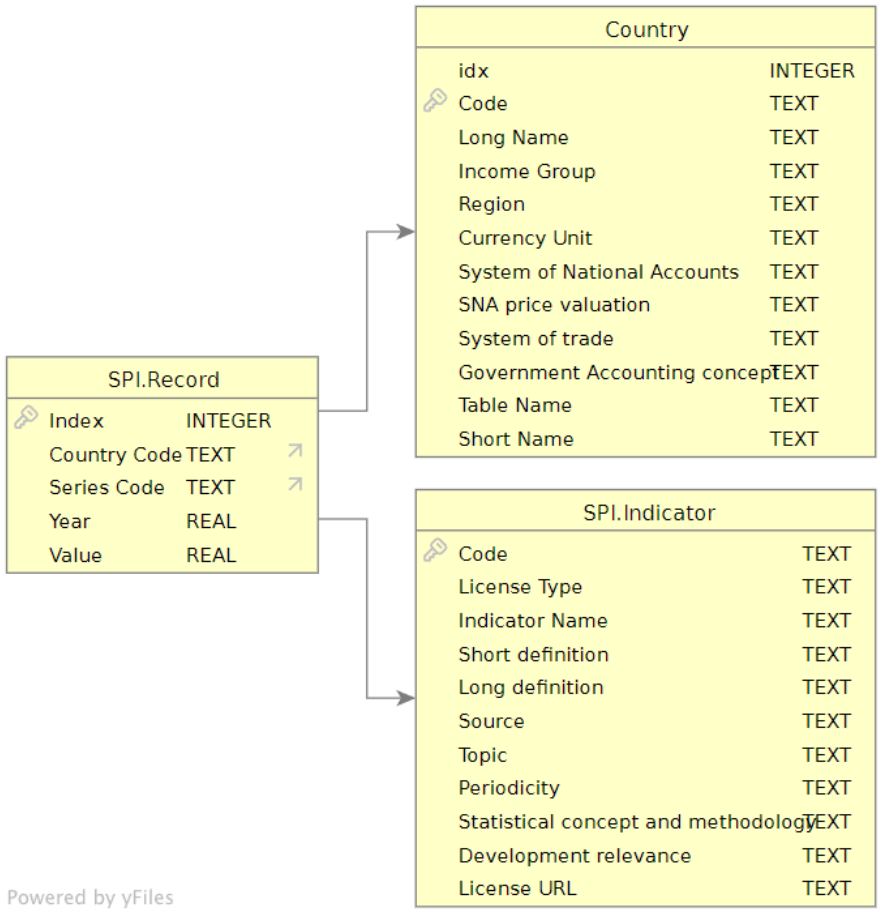


Figure 3. Entity Relationship Diagram of Database Tables

Dataset Extraction and Filtering

Once the data tables were created, Structured Query Language (SQL) views were created in order to group all the data onto a single table. Since most of the data was stored in identifier keys, several joins were used in order to ensure that the relevant field names were instead displayed. Once this view of all records was created, it was exported and ready for filtering.

The first filtering stage of the dataset required the handling of missing data. We performed two operations, first a column-wise drop and then a row-wise drop. Any columns which had a missing percentage of over 70% were dropped, then any rows which had a missing

percentage of over 90% were dropped. These percentages were chosen in order to minimize necessary imputation while also maximizing the number of observations in the dataset. The next filtering stage was a manual field selection to ensure that the columns related to at least one of the following categories: socioeconomic, environmental, or health. The full filtering process with dimension counts at each stage are represented in Figure 1, Data Pipeline.

Dataset Information

After applying the final filtering criteria, 171 countries were retained for analysis with the data spanning the years 2004 to 2021. Countries that did not have sufficient observations in any of the twenty years were excluded from the final analysis. The distribution of countries is detailed in Table 1 below, in which all countries are listed by their region and income group. By extending the data extraction to 2004, many countries and regions were able to be incorporated into the dataset. This diversity allows the study to capture a wide sample of geographic and socio-economic contexts.

Table 1. Countries by Region and Income Group

Region	Income	Count	Custom
Sub-Saharan Africa	Lower middle income	19	Angola, Benin, Cabo Verde, Cameroon, Comoros, Congo, Côte d'Ivoire, Eswatini, Ghana, Guinea, Kenya, Lesotho, Mauritania, Nigeria, Senegal, São Tomé and Príncipe, Tanzania, Zambia, Zimbabwe
Sub-Saharan Africa	Upper middle income	5	Botswana, Gabon, Mauritius, Namibia, South Africa
Sub-Saharan Africa	Low income	19	Burkina Faso, Burundi, Central African Republic, Chad, Dem. Rep. Congo, Ethiopia, Guinea-Bissau, Liberia, Madagascar, Malawi, Mali, Mozambique, Niger, Rwanda, Sierra Leone, Sudan, The Gambia, Togo, Uganda

South Asia	Low income	1	Afghanistan
South Asia	Lower middle income	6	Bangladesh, Bhutan, India, Nepal, Pakistan, Sri Lanka
South Asia	Upper middle income	1	Maldives
North America	High income	2	Canada, United States
Middle East & North Africa	Upper middle income	4	Algeria, Iran, Iraq, Libya
Middle East & North Africa	High income	8	Bahrain, Israel, Kuwait, Malta, Oman, Qatar, Saudi Arabia, United Arab Emirates
Middle East & North Africa	Lower middle income	6	Djibouti, Egypt, Jordan, Lebanon, Morocco, Tunisia
Middle East & North Africa	Low income	2	Syrian Arab Republic, Yemen
Latin America & Caribbean	Upper middle income	16	Argentina, Belize, Brazil, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Jamaica, Mexico, Paraguay, Peru, St. Lucia, St. Vincent and the Grenadines, Suriname
Latin America & Caribbean	High income	7	Barbados, Chile, Guyana, Panama, The Bahamas, Trinidad and Tobago, Uruguay
Latin America & Caribbean	Lower middle income	4	Bolivia, Haiti, Honduras, Nicaragua
Latin America & Caribbean	Low income	1	Venezuela
Europe & Central Asia	Upper middle income	13	Albania, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Georgia, Kazakhstan, Moldova, Montenegro, North Macedonia, Serbia, Türkiye, Ukraine
Europe & Central Asia	High income	31	Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Portugal, Romania, Russia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, United Kingdom
Europe & Central Asia	Lower middle income	3	Kyrgyz Republic, Tajikistan, Uzbekistan

East Asia & Pacific	High income	6	Australia, Brunei, Japan, Korea, New Zealand, Singapore
East Asia & Pacific	Lower middle income	10	Cambodia, Lao PDR, Myanmar, Papua New Guinea, Philippines, Samoa, Solomon Islands, Timor-Leste, Vanuatu, Viet Nam
East Asia & Pacific	Upper middle income	7	China, Fiji, Indonesia, Malaysia, Mongolia, Thailand, Tonga

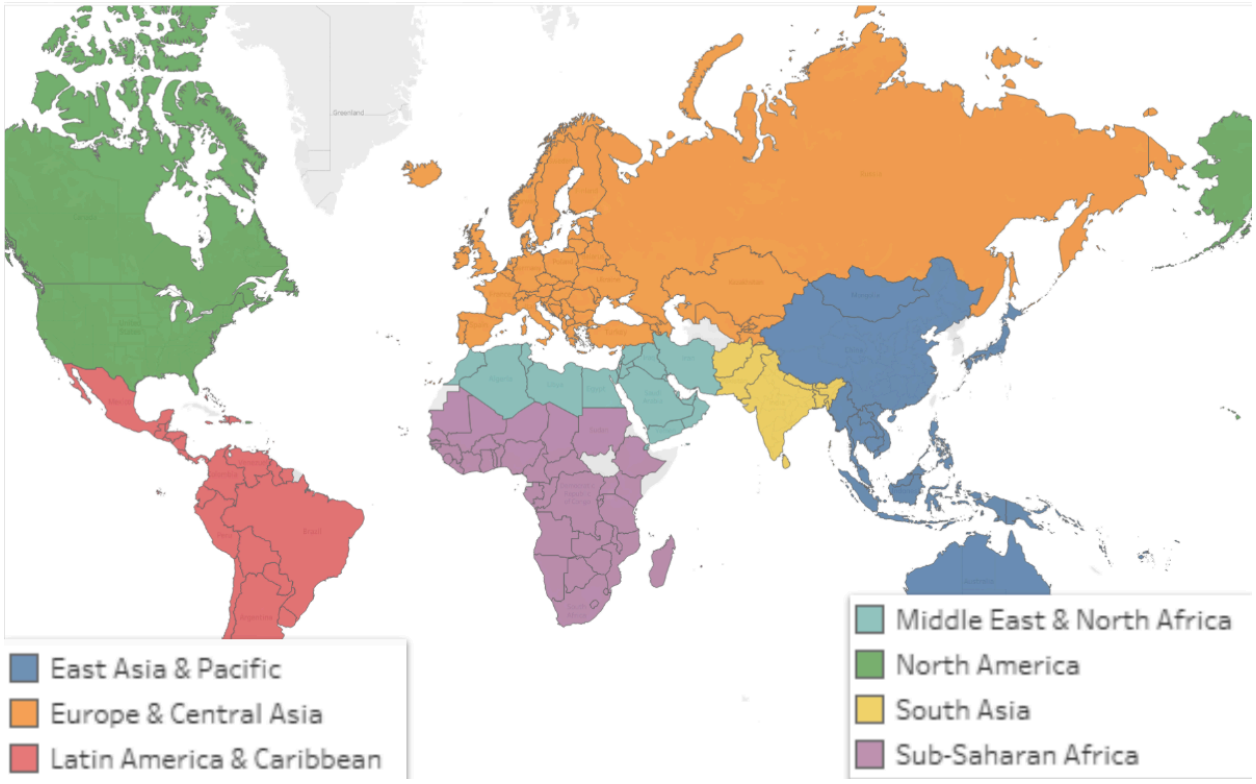


Figure 4. Countries in Analysis by Region

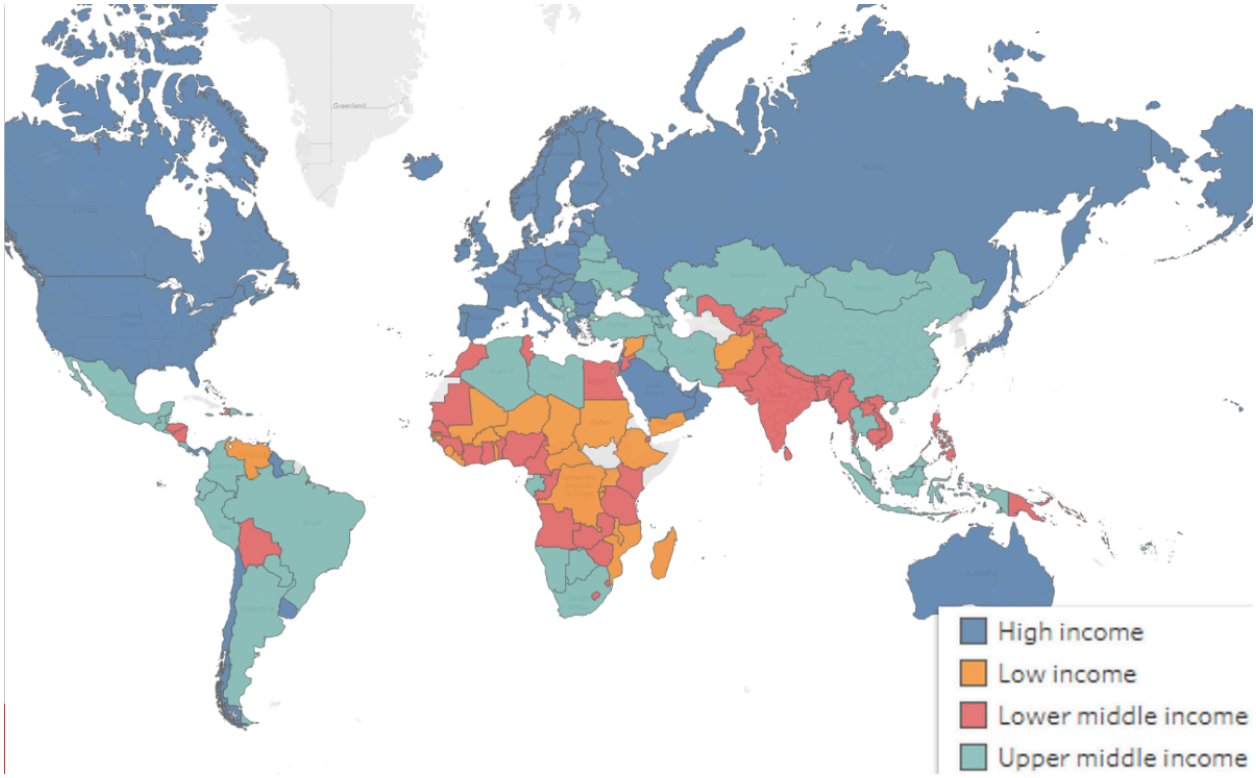


Figure 5. Countries in Analysis by Income Group

A grand total of 2,886 observations remained after the final filtering process. To ensure data quality and reliability, an observation for a given country was removed if more than 90% of the data within that observation were missing. This strict threshold was chosen to ensure minimal data imputation would be required. Dataset completeness was examined over the years, revealing a significant decline after 2021. Consequently, observations from the years 2022 to 2024 were entirely dropped from the analysis, as these years did not meet the minimum data completeness criteria required for the filter. Finally, after the manual selection of relevant fields was completed, a total of 286 indicators remained in the dataset.

Table 2. Observation Count by Year

Year	East Asia & Pacific	Europe & Central Asia	Latin America & Caribbean	Middle East & North Africa	North America	South Asia	Sub-Saharan Africa	Grand Total
2004	21	44	25	14	2	5	33	**144**
2005	20	45	28	18	2	7	38	**158**
2006	21	45	28	18	2	7	38	**159**
2007	21	46	28	18	2	7	38	**160**
2008	21	46	28	17	2	8	39	**161**
2009	22	46	28	18	2	8	41	**165**
2010	23	47	28	18	2	8	42	**168**
2011	22	47	28	18	2	8	42	**167**
2012	23	47	28	18	2	8	41	**167**
2013	23	47	28	19	2	8	41	**168**
2014	23	47	27	19	2	8	40	**166**
2015	23	47	27	18	2	8	42	**167**
2016	23	47	27	19	2	8	42	**168**
2017	23	47	27	18	2	8	43	**168**
2018	23	47	26	16	2	8	41	**163**
2019	23	47	26	18	2	8	41	**165**
2020	21	47	26	17	2	8	39	**160**
2021	12	44	18	8		5	25	**112**

Table 3. Indicator Category Counts Before Stepwise Selection.

Indicator Category	Count of Indicators
Environmental	81
Health	76
Socioeconomic	129
Grand Total	**286**

Dataset Preparation

All remaining missing values in the dataset were imputed using a K-Nearest Neighbors (KNN) imputer, a method chosen over simpler techniques (such as mean imputation) to better preserve the distinctive characteristics of each country. Next, to handle outliers, winsorization

was applied by capping values at 3.5 standard deviations from the mean. This method was used in order to mitigate outlier impact while still retaining the overall data distribution. Finally, each column was standardized using a standard scaler, removing the mean and rescaling the data so that the standard deviation of each feature is equal to one.

Dataset Splitting

To assess the distinct contributions of socioeconomic, environmental, and health indicators in predicting life expectancy at birth, the dataset was partitioned into three separate indicator arrays, each only containing columns which belong exclusively to one of the indicator categories. This partitioning enabled us to evaluate the predictive capability and relative importance of features within each domain independently. In addition, a model based on the full, unpartitioned dataset was constructed and evaluated to compare its performance against the category-specific models.

Feature Selection

To improve the interpretability and stability of the prediction models, we implemented a stepwise feature selection procedure designed to reduce dimensionality and to mitigate multicollinearity. First, the variance inflation factor (VIF) was calculated for each predictor variable within each category-specific dataset. In our analysis, a threshold VIF value of 4 was selected. Any predictor exhibiting a VIF above this threshold was considered excessively collinear with other predictors and thus subject to removal. The following stepwise elimination process was applied: at each iteration, the predictor with the highest VIF value (provided it exceeded the threshold of 4) was removed from the model. Following each removal, VIF calculations were repeated on the reduced set of predictors. This iterative process continued until all remaining variables in the model exhibited VIF values at or below 4. Finally, a preliminary

OLS model was fit and tested, and any indicators which had a p-value greater than 0.05 were manually identified and dropped. This approach ensured that the final set of predictors possessed a minimal degree of redundancy.

Analysis Methods

The analysis framework, as shown in Figure 2, Analysis Pipeline, primarily uses two methods: Ordinary Least Squares Regression (OLS) and K-means clustering.

Ordinary Least Squares Regression

A stepwise Ordinary Least Squares (OLS) regression was conducted on each of the three indicator sets—environmental, socioeconomic, and health—to identify the most influential predictors of life expectancy. OLS was selected primarily for model simplicity and interpretability, allowing for the clear quantification of relationships between independent variables and the dependent variable (life expectancy). Stepwise selection was utilized through Variance Inflation Factor (VIF) thresholds to mitigate multicollinearity and high dimensionality of the initial dataset. Last, the extracted p-values were verified to ensure the inclusion of statistically significant predictors. By refining the model in this way, we could isolate the most critical factors shaping life expectancy and gauge their relative importance. Ultimately, this method aimed to answer which indicators within each domain are key drivers of life expectancy, as well as the extent to which changes in these indicators translate into measurable shifts in health outcomes across different countries.

K-Means Clustering

After the regression analysis, k-means clustering was performed at five-year intervals beginning in 2005 to capture the changes of country-level groupings based on their collective indicator profiles. K-means was selected for its simplicity and effectiveness in revealing

underlying data structures without predefined group labels. For each category (environmental, socioeconomic, and health), the variables used for clustering were those that both met the Variance Inflation Factor (VIF) threshold and demonstrated p-value significance in the OLS model. This approach aims to uncover how groups of countries sharing similar characteristics progress or diverge over time, and whether these shifts help explain variations in life expectancy.

Results

Exploratory Data Analysis

Exploratory data analysis (EDA) is the first step in working with our clean and prepared dataset. It helps us understand the dataset, suggest hypotheses related to the drivers of life expectancy at birth, and assess assumptions on which statistical inference will be based.

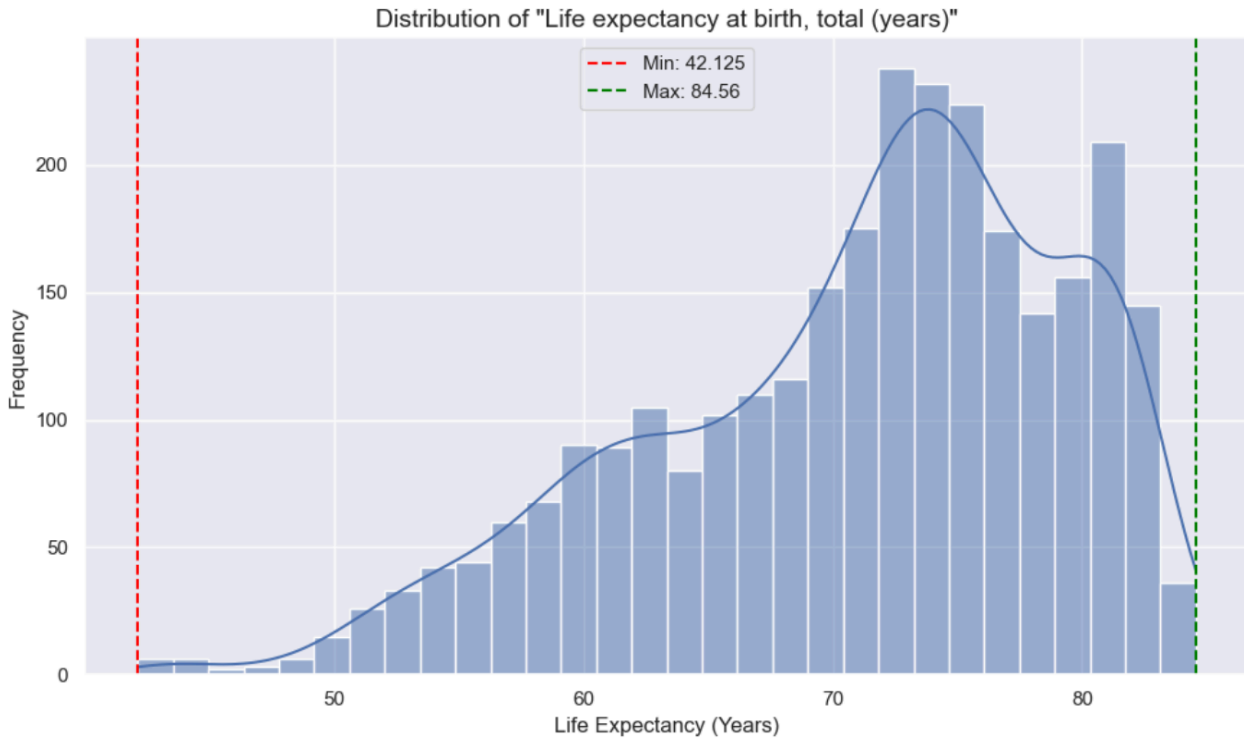


Figure 6. Distribution of Target Variable (LEB)

Figure 6's histogram shows that most countries cluster between 70 and 80 years of life expectancy, with the overall mean at 70.9 and the median slightly higher at 72.6. The left-skew indicates a longer tail of low-life-expectancy outliers, with a few extreme values below 50 years. The interquartile range (roughly 68 to 76 years) captures the bulk of the data.

Country	2004	2009	2014	2019	2021	Percent Change (2004-2021)
Ethiopia	53.24	58.8	62.9	65.84	64.97	22.05%
Angola	50.59	55.75	60.04	62.45	61.64	21.84%
Uganda	51.55	56.49	60.41	62.99	62.7	21.65%
Lesotho	43.71	44.03	50.03	54.17	53.06	21.39%
Tanzania	55.6	58.89	63.87	66.99	66.2	19.08%
Rwanda	55.49	61.96	64.94	66.44	66.07	19.07%
Botswana	51.62	58.28	62.61	65.46	61.14	18.45%
Niger	52.59	57.4	60.79	62.9	61.58	17.09%
Namibia	51.25	55.5	59.82	63.08	59.27	15.66%
South Africa	54.04	57.45	63.38	66.18	62.34	15.35%

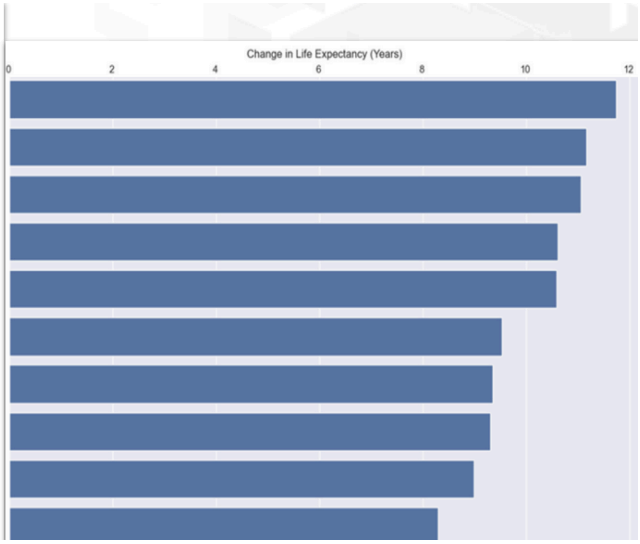
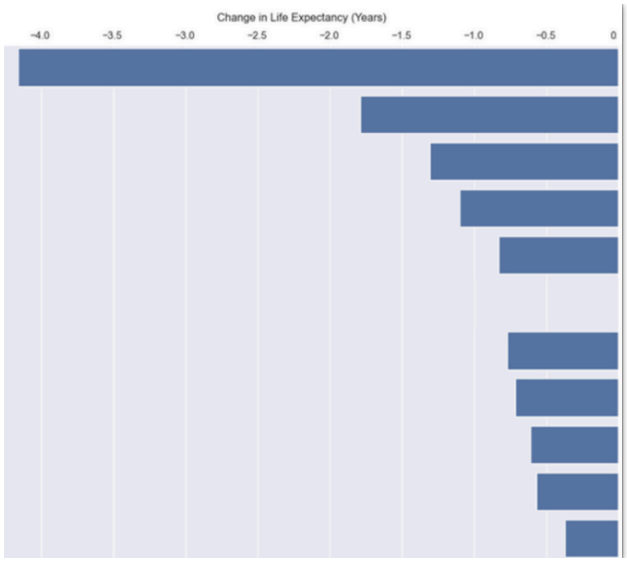


Figure 7. Countries with Most Positive Net Change in LEB



Country	2004	2009	2014	2019	2021	Percent Change (2004-2021)
Mexico	74.37	74.12	74.8	74.2	70.21	-5.59%
Oman	74.33	74.68	77.45	78	72.54	-2.41%
Costa Rica	78.33	78.67	78.77	79.43	77.02	-1.67%
Bulgaria	72.56	73.41	74.47	75.11	71.46	-1.52%
Philippines	70.04	70.63	71.15	71.86	69.27	-1.10%
Bosnia and Herzegovina	76.13	76.98	75.66	77.24	75.3	-1.09%
Tunisia	74.49	75.27	75.65	75.99	73.77	-0.96%
Ecuador	74.28	75.18	76.62	77.3	73.67	-0.82%
Uruguay	76.01	76.74	77.37	77.51	75.44	-0.75%
Bolivia	64	65.88	67.16	67.84	63.63	-0.58%

Figure 8. Countries with Most Negative Net Change in LEB

Figure 7 above highlights the countries with the largest gains in life expectancy at birth. All of the top gainers, each improving by more than 15%, are located in sub-Saharan Africa, underscoring the importance of regional metadata in our dataset. In comparison, the steepest declines in Figure 8 were relatively small, with Mexico having the largest drop at just under 6%.

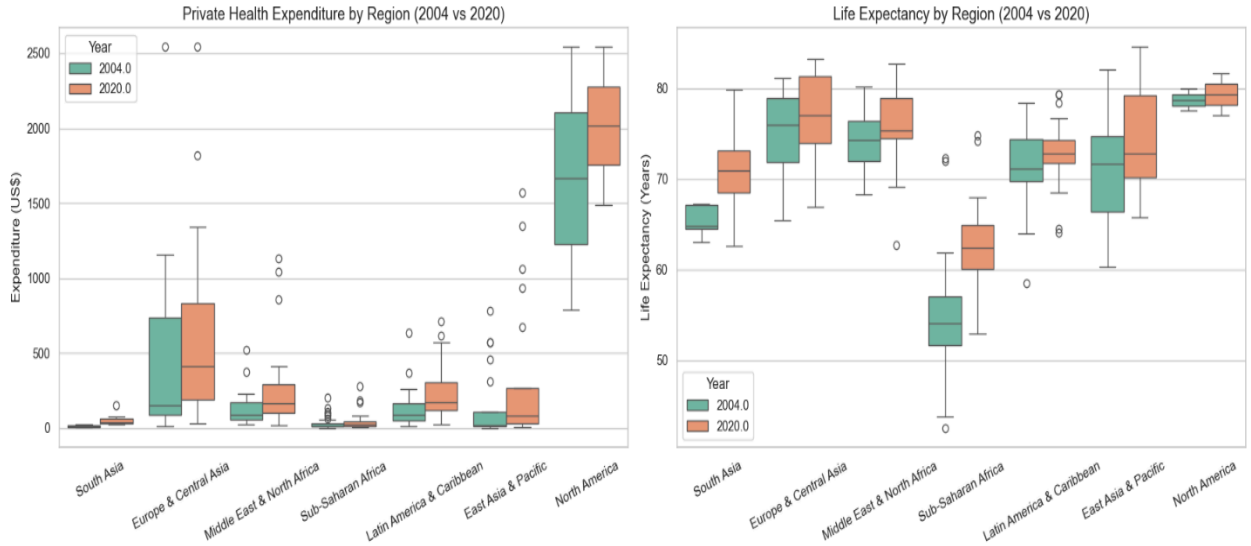


Figure 9. Health Expenditure per Capita and LEB by Region in 2004 and 2020

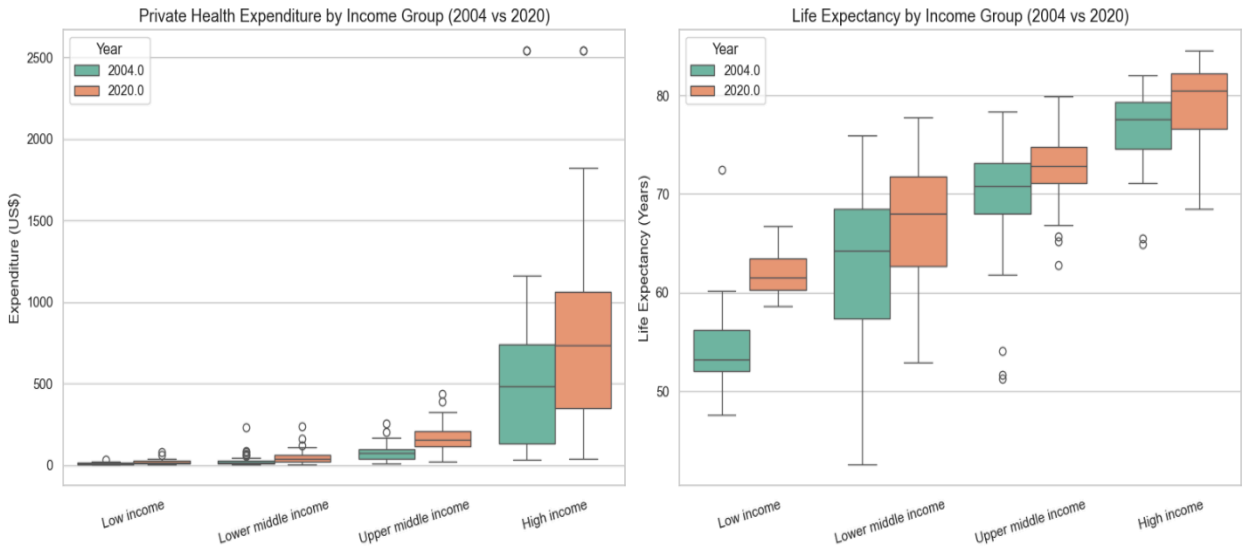


Figure 10. Health Expenditure per Capita and LEB by Income Group in 2004 and 2020

Figures 9 and 10 above provide an overview of the relationship between private health expenditure per capita and life expectancy at birth (LEB) across different regions and income groups from 2004 and 2020. Figure 9 illustrates the variations in private health expenditure and life expectancy by region, highlighting notable differences in health outcomes across regions. For instance, North America exhibits the highest private health expenditure per capita, corresponding with a relatively high life expectancy. Sub-Saharan Africa, however, shows much lower health expenditures and life expectancy, suggesting a correlation between economic investment in healthcare and health outcomes. Figure 10 focuses on the disparities across income groups, revealing the intuitive pattern in which higher income groups experience both higher health expenditures and longer life expectancy. These figures emphasize that although there is a link between health expenditure and life expectancy, there are many more predictors at play.

Prediction Models Results

Using the three indicator arrays, three separate ordinary least squares (OLS) regression models were developed to serve as baseline predictions and to evaluate the relative importance of predictors across socioeconomic, environmental, and health domains. In each model, "Life expectancy at birth, total (years)" was configured as the dependent variable. This approach allowed us to isolate and determine the explanatory power of each group of indicators independently, as well as to compare their contributions toward predicting life expectancy. The final models were evaluated using key fit statistics including R-squared, adjusted R-squared, root mean squared error on the test set, the F-statistic along with its associated p-value, as well as model selection criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These metrics are listed in Table 4 on the next page.

The health predictors model achieved an R-squared of 0.881, indicating that approximately 88.1% of the variance in life expectancy is explained by health indicators. One contributing factor of this high value is that the life expectancy at birth in the dataset is derived from period life tables (“Metadata Glossary”). Period life tables provide a snapshot of a population’s current mortality patterns, meaning that they reflect the immediate health conditions and risks present at the time of data collection. As a result, measures such as maternal mortality, disease incidence, and healthcare expenditure are closely aligned with the mortality patterns observed in period life tables.

Because these health indicators are contemporaneous with the mortality data, they naturally demonstrate a strong association with life expectancy outcomes. Essentially, the period life table serves as a “real-time” summary of current health and mortality conditions; hence, any variable that effectively captures the current state of healthcare, nutrition, and disease burden can

significantly improve the predictive power of the model. This inherent connection between the measurement method and the health indicators partially explains the high R-squared value observed in our health predictors model. However, the socioeconomic (R-squared = 0.806) and environmental (R-squared = 0.665) models also demonstrated meaningful predictive capabilities.

Table 4. OLS Model Fit Comparison

Statistic	Environmental Model	Socioeconomic Model	Health Model
Dependent Variable	Life expectancy at birth, total (years)	Life expectancy at birth, total (years)	Life expectancy at birth, total (years)
Model Type	Ordinary Least Squares (OLS)	Ordinary Least Squares (OLS)	Ordinary Least Squares (OLS)
Observations (n)	2,308	2,308	2,308
R-squared	0.665	0.806	0.881
Adjusted R-squared	0.662	0.805	0.881
Test RMSE	4.89	3.929	2.988
F-statistic	252.4	453.3	1,704
Prob (F-statistic)	< 0.001	< 0.001	< 0.001
Log-Likelihood	-6,912.80	-6,280.10	-5,716.30
AIC	13,860	12,600	11,450
BIC	13,970	12,730	11,520
Degrees of Freedom (Model)	18	21	10
Degrees of Freedom (Residuals)	2,289	2,286	2,297

Table 5. OLS Coefficients - Environmental Model

Variable	Coefficient	Std Err	t	P-value	95% CI
const	70.9735	0.101	702.074	0	[70.775, 71.172]
Carbon dioxide (CO2) emissions	2.7249	0.143	19.106	0	[2.445, 3.005]
Fertilizer consumption	1.4113	0.118	11.939	0	[1.180, 1.643]
Water productivity, total	1.0815	0.122	8.884	0	[0.843, 1.320]

Renewable internal freshwater resources per capita	0.8305	0.122	6.788	0	[0.591, 1.070]
GOAL 13: Climate Action (5 year moving average)	0.5918	0.109	5.415	0	[0.377, 0.806]
Capture fisheries production	0.5837	0.147	3.978	0	[0.296, 0.871]
Population density	0.5699	0.122	4.653	0	[0.330, 0.810]
Forest area	0.3494	0.124	2.829	0.005	[0.107, 0.592]
Aquaculture production	0.3306	0.153	2.154	0.031	[0.030, 0.632]
Arable land	0.3142	0.128	2.45	0.014	[0.063, 0.566]
Coal rents	-0.2533	0.109	-2.317	0.021	[-0.468, -0.039]
Mineral rents	-0.4826	0.115	-4.211	0	[-0.707, -0.258]
Nitrous oxide (N2O) emissions (total)	-0.5	0.123	-4.05	0	[-0.742, -0.258]
Natural gas rents	-0.6795	0.118	-5.761	0	[-0.911, -0.448]
Oil rents	-1.0938	0.123	-8.915	0	[-1.334, -0.853]
Methane (CH4) emissions from Building	-1.1526	0.144	-8.005	0	[-1.435, -0.870]
Rural population growth	-1.2344	0.122	-10.091	0	[-1.474, -0.995]
Forest rents	-3.0727	0.124	-24.788	0	[-3.316, -2.830]

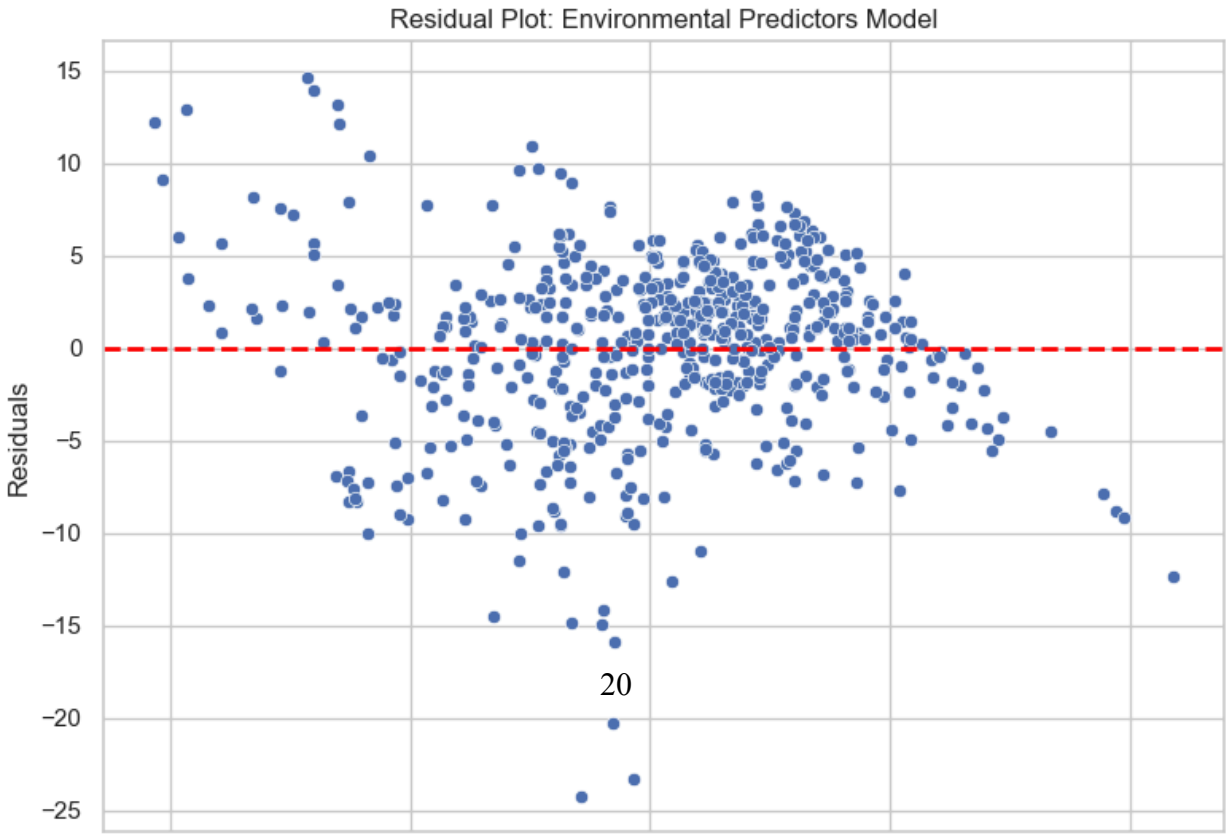


Table 6. OLS Coefficients - Socioeconomic Model

Variable	Coefficient	Std Err	t	P-value	95% CI
const	70.9735	0.077	922.895	0	[70.823, 71.124]
Agriculture, forestry, and fishing, value added per worker	1.6917	0.097	17.463	0	[1.502, 1.882]
External balance on goods and services	1.2373	0.094	13.139	0	[1.053, 1.422]
Commercial bank branches	1.1148	0.099	11.313	0	[0.922, 1.308]
Claims on central government,	0.909	0.087	10.433	0	[0.738, 1.080]
International tourism, expenditures for travel items	0.7047	0.115	6.148	0	[0.480, 0.930]
GOAL 11: Sustainable Cities and Communities	0.5353	0.082	6.494	0	[0.374, 0.697]
Ores and metals imports	0.3954	0.09	4.412	0	[0.220, 0.571]
International tourism, number of arrivals	0.3202	0.108	2.964	0.003	[0.108, 0.532]
Income Group_Upper middle income	0.3141	0.091	3.456	0.001	[0.136, 0.492]
Travel services	-0.1723	0.083	-2.064	0.039	[-0.336, -0.009]
Unemployment, youth total (% of total labor force ages 15-24)	-0.2127	0.09	-2.355	0.019	[-0.390, -0.036]
Communications, computer, etc.	-0.2219	0.084	-2.63	0.009	[-0.387, -0.056]
Final consumption expenditure	-0.2576	0.081	-3.169	0.002	[-0.417, -0.098]

Adjusted savings: energy depletion	-0.3507	0.095	-3.695	0	[-0.537, -0.165]
Region_South Asia	-0.3778	0.092	-4.122	0	[-0.557, -0.198]
Region_East Asia & Pacific	-0.4308	0.093	-4.639	0	[-0.613, -0.249]
Region_Latin America & Caribbean	-0.4484	0.1	-4.503	0	[-0.644, -0.253]
Income Group_Low income	-0.4992	0.097	-5.153	0	[-0.689, -0.309]
Inflation, GDP deflator	-0.7417	0.081	-9.205	0	[-0.900, -0.584]
Primary education, pupils	-0.949	0.104	-9.154	0	[-1.152, -0.746]
Region_Sub-Saharan Africa	-4.6515	0.118	-39.512	0	[-4.882, -4.421]

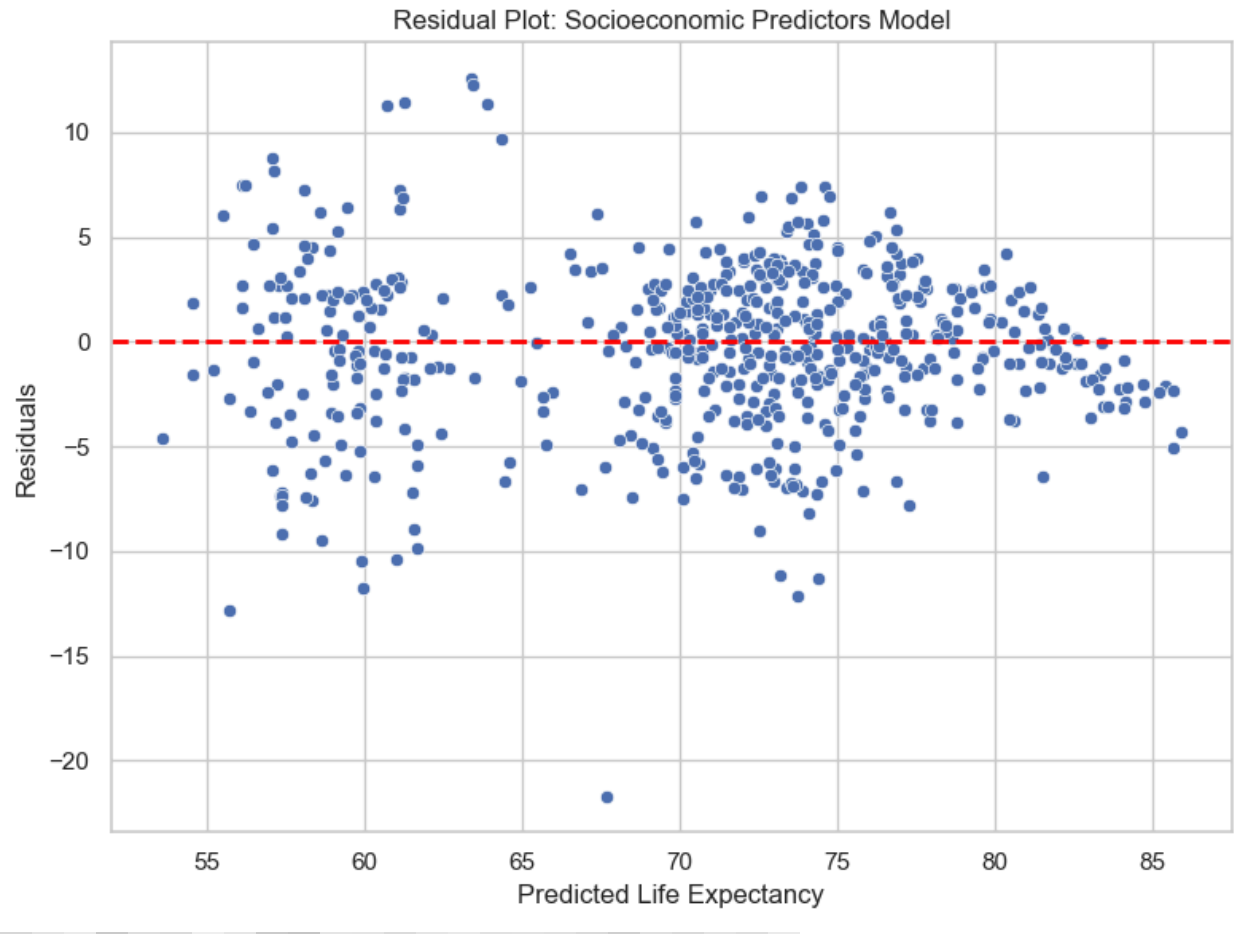


Table 7. OLS Coefficients - Health Model

Variable	Coefficient	Std Err	t	P-value	95% CI
const	70.9735	0.06	1181.082	0	[70.856, 71.091]
Domestic private health expenditure per capita	1.8903	0.077	24.432	0	[1.739, 2.042]
Number of maternal deaths	0.5333	0.108	4.952	0	[0.322, 0.744]
Population ages 0-14, total	-0.2172	0.096	-2.257	0.024	[-0.406, -0.028]
Prevalence of overweight	-0.3456	0.069	-5.004	0	[-0.481, -0.210]

Total alcohol consumption per capita	-0.7015	0.073	-9.653	0	[-0.844, -0.559]
People practicing open defecation	-0.7328	0.082	-8.893	0	[-0.894, -0.571]
Prevalence of undernourishment	-0.8234	0.085	-9.702	0	[-0.990, -0.657]
Incidence of tuberculosis	-2.2119	0.077	-28.904	0	[-2.362, -2.062]
Lifetime risk of maternal death	-4.1798	0.095	-44.107	0	[-4.366, -3.994]

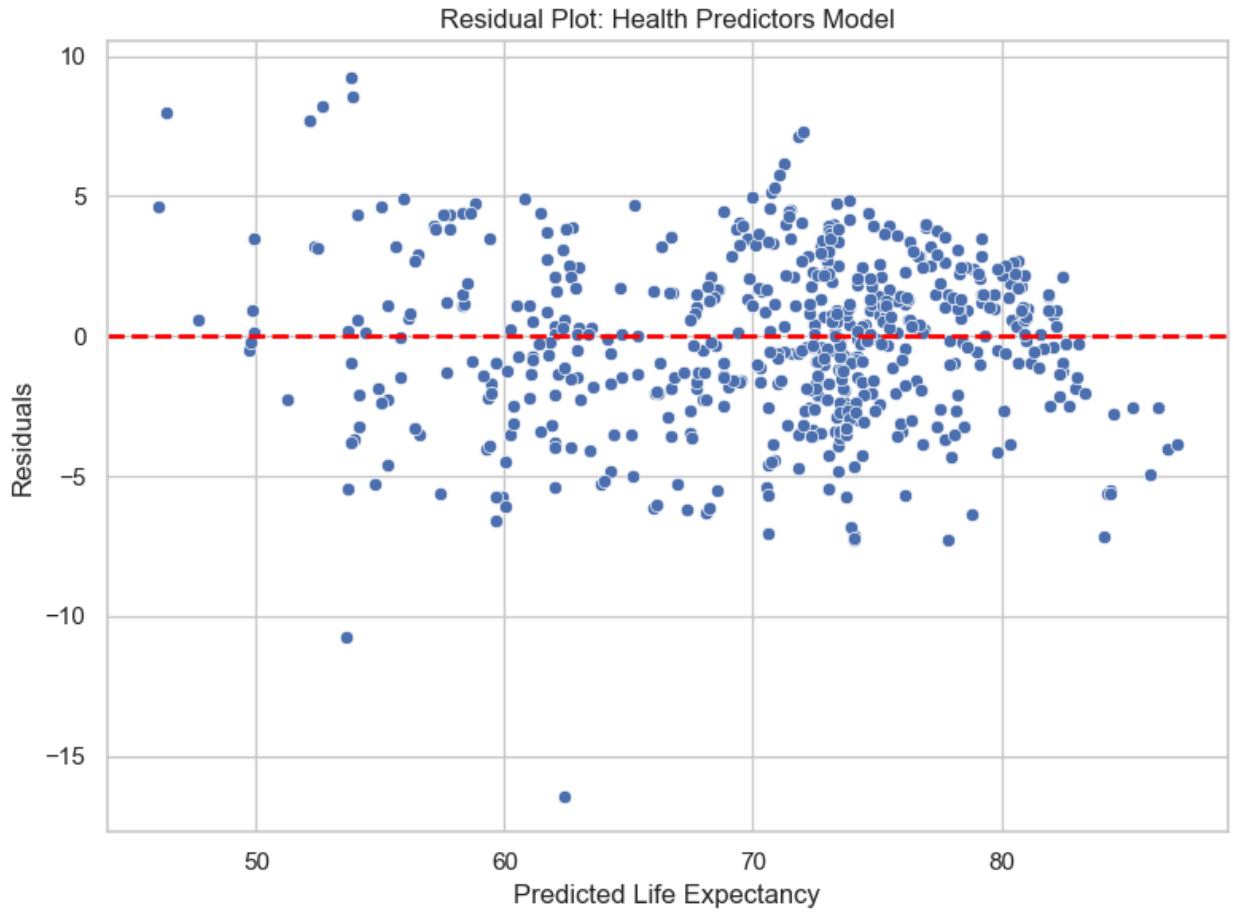


Figure 13. Health OLS Model Residuals

Parametric and Non-Parametric Models

Another step in the prediction framework was to change the model to compare and contrast performance. The prediction models had similar performance characteristics across each of the three datasets.

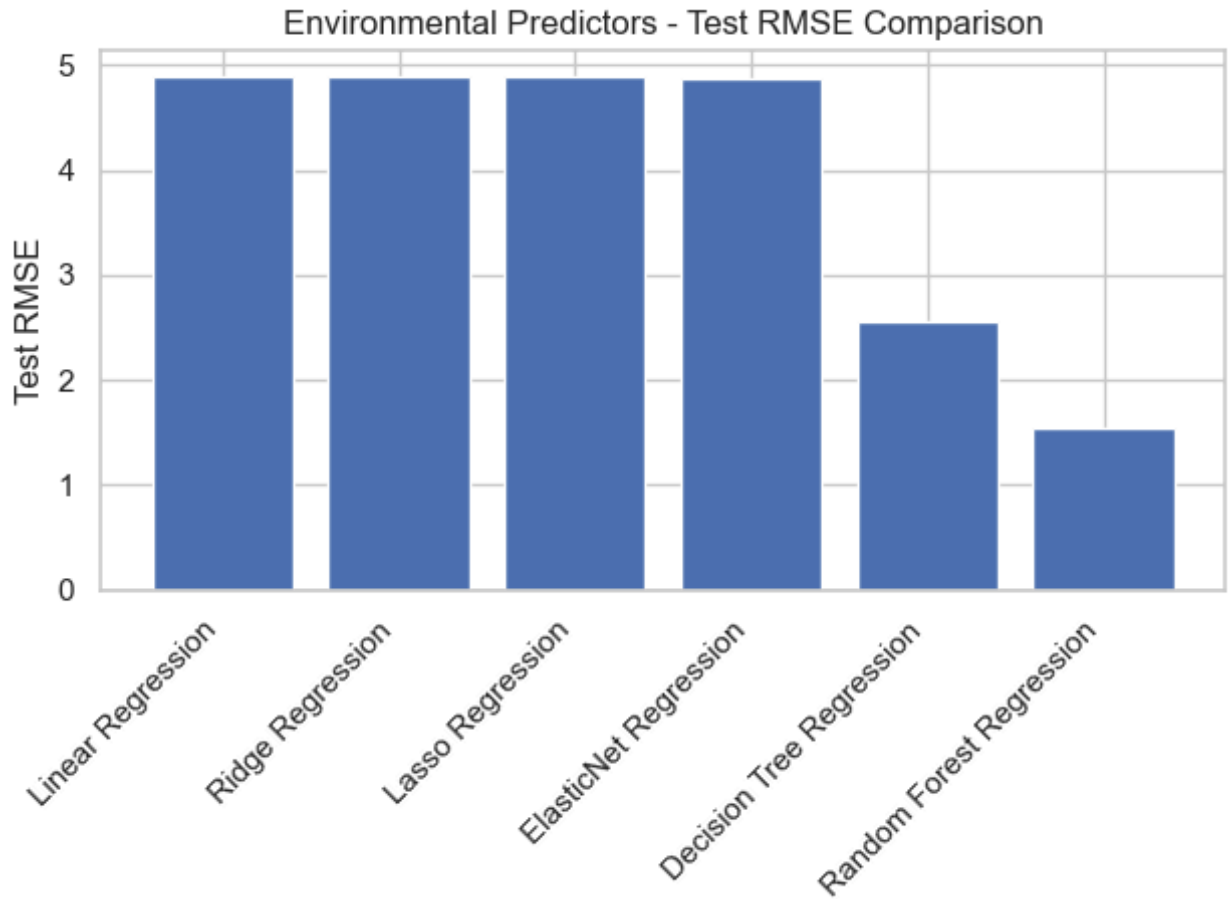


Figure 14. Parametric vs. Non-Parametric Model Performance

Each parametric model performed similarly in terms of RMSE. The classical decision tree performed better, and the random forest outperformed even the decision tree. This is likely because the non-parametric models could capture more complex relationships in the data, which parametric models were unable to do. While the parametric models (Linear, Ridge, Lasso, and

Elastic Net) showed similar results in terms of RMSE, they maintain advantages in terms of simplicity and extraction of statistical inference.

Cluster Analysis

Given the dataset, our analysis sought questions related to the indicators themselves: What insights emerge when countries are clustered based on environmental, socioeconomic, and health indicators? Do countries with similar indicator profiles follow comparable life expectancy trends? And how do these indicators change over time? To investigate these questions, we conducted a k-means clustering analysis for each indicator category. Using the indicator columns which remained after the VIF stepwise selection, the baseline cluster labels for each country within each indicator group were created for the baseline year 2005. These baseline clusters were then tracked over subsequent five-year intervals (2010, 2015, and 2020) to observe the movement of cluster centroids over time. By comparing these temporal changes to patterns in life expectancy outcomes, we aim to reveal how evolving national profiles relate to population health over time.

Clustering by Environmental Indicators

For the environmental indicators, our clustering analysis (see Table 8 on the next page) yielded a two-cluster solution. The ANOVA results confirmed that the variables GOAL 13: Climate Action score ($F = 156$, $p < 0.001$), CO_2 emissions per capita excluding LULUCF ($F = 13.14$, $p < 0.001$), forest area ($F = 4.767$, $p = 0.0305$), and water productivity ($F = 4.456$, $p = 0.03638$) significantly distinguished the two clusters. Out of 158 observations, Cluster 1 (75 members) and Cluster 2 (83 members) displayed distinct environmental profiles. In addition, Table 9 (pg.) presents the change over time of these significant environmental indicators. In Cluster 1, average life expectancy increased from 69.54 years in 2005 to 73.11 years in 2020,

while key environmental measures such as water productivity experienced a 47.00% increase. In contrast, Cluster 2, although showing a lower baseline in life expectancy (68.24 years in 2005), exhibited different patterns in variables like GOAL 13 scores and CO_2 emissions. These findings suggest that environmental factors are not only significant in differentiating country groups but also evolve distinctly over time, potentially influencing long-term health outcomes.

Table 8. K-Means Clustering by Environmental Predictors: Analysis of Variance

Variable	F-statistic	p-value
Avg. GOAL 13: Climate Action (5 year moving average)	156	0
Avg. Carbon dioxide (CO2) emissions	13.14	0.0003899
Avg. Forest area	4.767	0.0305
Avg. Water productivity, total	4.456	0.03638
Avg. Oil rents	2.797	0.09646
Avg. Arable land	2.543	0.1128
Avg. Natural gas rents	2.517	0.1147
Avg. Nitrous oxide (N2O)	1.994	0.16
Avg. Population density	1.513	0.2205
Avg. Coal	1.036	0.3103
Avg. Aquaculture production	0.7616	0.3842
Avg. Methane (CH4) emissions from Building	0.6843	0.4094
Avg. Mineral rents	0.6706	0.4141
Avg. Fertilizer consumption	0.5655	0.4532
Avg. Rural population growth	0.5248	0.4699
Avg. Renewable internal freshwater resources per capita	0.1652	0.685
Avg. Forest rents	0.001274	0.9716
Avg. Capture fisheries production	6.32E-05	0.9937

Number of Clusters: 2, **Number of Points**: 158, **Between-group Sum of Squares**: 40.996, **Within-group Sum of Square**s: 85.601, **Total Sum of Squares**: 126.6, **Cluster 1**: 75, **Cluster 2**: 83

Table 9. Environmental Predictors Over Time

Indicator Averages by Year	2005	2010	2015	2020	Delta
Environment - Cluster 1					
Average of Life expectancy at birth, total (years)	69.54	71.12	72.74	73.11	5.12%
Average of GOAL 13: Climate Action	0.00	0.17	0.32	0.56	
Average of Carbon dioxide (CO2) emissions	6.60	6.55	6.14	5.85	11.32 %
Average of Forest area	27.62	27.60	27.95	26.85	-2.80%
Average of Water productivity, total	59.34	64.84	82.77	87.22	47.00 %
Environment - Cluster 2					
Average of Life expectancy at birth, total (years)	68.24	70.15	71.67	71.92	5.38%
Average of GOAL 13: Climate Action	1.00	1.00	0.67	0.81	18.83 %
Average of Carbon dioxide (CO2) emissions	3.22	3.34	3.20	2.98	-7.53%
Average of Forest area	36.10	35.96	35.54	35.86	-0.68%
Average of Water productivity, total	32.25	39.51	44.43	45.64	41.53 %

Clustering by Socioeconomic Indicators

The cluster analysis by the socioeconomic indicators produced a three-cluster solution (see Table 10 on the next page). The ANOVA diagnostics revealed that variables such as "Income Group Upper middle income" ($F = 77.5$, $p < 0.001$), "Region Sub-Saharan Africa" ($F = 65.6$, $p < 0.001$), and "Commercial bank branches (per 100,000 adults)" ($F = 21.15$, $p < 0.001$) were key differentiators among clusters. Table 11 (pg.) tracks the averages of the numerical socioeconomic indicators over the five-year intervals. demonstrating that clusters exhibit distinct temporal trends. For example, one cluster (Cluster 1) displays relatively high life expectancy and consistent economic advantages, while another (Cluster 2) shows considerably lower life expectancy coupled with markedly lower banking and agricultural performance, and Cluster 3

falls in between. The evolution of these clusters provides insights into how economic factors might be underlying divergent health outcomes across countries.

Table 10. K-Means Clustering by Socioeconomic Predictors: Analysis of Variance

Variable	F-statistic	p-value
Avg. Income Group Upper middle income	77.5	0
Avg. Region Sub-Saharan Africa	65.6	0
Avg. Income Group Low income	27.97	4.25E-11
Avg. Commercial bank branches (per 100,000 adults)	21.15	7.57E-09
Avg. Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)	11.99	1.44E-05
Avg. Ores and metals imports (% of merchandise imports)	9.485	0.00013
Avg. Region Latin America & Caribbean	8.861	0.0002272
Avg. Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	6.75	0.001547
Avg. International tourism, expenditures for travel items (current US$)	5.839	0.003591
Avg. Inflation, GDP deflator (annual %)	5.614	0.004429
Avg. International tourism, number of arrivals	3.827	0.02386
Avg. Communications, computer, etc. (% of service exports, BoP)	3.06	0.04975
Avg. External balance on goods and services (% of GDP)	2.855	0.06058
Avg. Region South Asia	1.819	0.1656
Avg. Travel services (% of commercial service imports)	1.732	0.1804
Avg. GOAL 11: Sustainable Cities and Communities (5 year moving average)	1.265	0.2851
Avg. Adjusted savings: energy depletion (% of GNI)	1.248	0.29
Avg. Primary education, pupils	1.04	0.3558
Avg. Claims on central government, etc. (% GDP)	0.5608	0.5719
Avg. Final consumption expenditure (annual % growth)	0.09264	0.9116

Number of Clusters: 3, **Number of Points**: 158, **Between-group Sum of Squares**: 70.74, **Within-group Sum of Squares**: 125.32, **Total Sum of Squares**: 196.06, **Cluster 1**: 81, **Cluster 2**: 33, **Cluster 3**: 44

Table 11. Socioeconomic Predictors Over Time

Indicator Averages by Year	2005	2010	2015	2020	Delta
Socioeconomic - Cluster 1					
Average of Life expectancy at birth, total (years)	73.897	75.066	76.386	76.708	3.80%
Average of Commercial bank branches (per 100,000 adults)	24.199	24.273	21.620	19.596	-19.02%
Average of Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)	18862.370	21723.341	25217.292	27633.866	46.50%
Average of Ores and metals imports (% of merchandise imports)	2.780	3.207	2.815	3.103	11.62%
Average of Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	15.991	17.577	17.768	17.791	11.26%
Socioeconomic - Cluster 2					
Average of Life expectancy at birth, total (years)	54.926	58.211	61.018	62.371	13.56%
Average of Commercial bank branches (per 100,000 adults)	2.244	4.025	5.212	6.326	181.87 %
Average of Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)	1365.428	1557.895	1751.192	2019.148	47.88%
Average of Ores and metals imports (% of merchandise imports)	1.297	1.391	1.691	1.318	1.59%
Average of Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	12.524	12.502	12.239	13.287	6.09%

Socioeconomic - Cluster 3					
Average of Life expectancy at birth, total (years)	69.777	71.479	72.799	72.574	4.01%
Average of Commercial bank branches (per 100,000 adults)	14.101	15.830	17.801	15.478	9.77%
Average of Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)	7845.183	8292.586	9137.110	10021.954	27.75%
Average of Ores and metals imports (% of merchandise imports)	2.034	2.291	2.607	2.810	38.14%
Average of Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	21.873	22.064	21.795	23.400	6.98%

Clustering by Health Indicators

For health indicators, the clustering analysis (summarized in Table 12 below) returned a two-cluster solution that significantly separated countries based on health metrics. All health indicators were effective differentiators among the clusters. In this case, Cluster 1 (48 observations) and Cluster 2 (110 observations) were defined by sharp differences across these critical variables. This clustering provides a clear distinction between countries with generally better health outcomes and those facing significant public health challenges.

Table 12. K-Means Clustering by Health Predictors: Analysis of Variance

Variable	F-statistic	p-value
Avg. Lifetime risk of maternal death (%)	96.6	0
Avg. People practicing open defecation (% of population)	90.56	0
Avg. Prevalence of undernourishment (% of population)	75.11	5.22E-15

Avg. Incidence of tuberculosis (per 100,000 people)	64.89	1.92E-13
Avg. Number of maternal deaths	35.66	1.53E-08
Avg. Prevalence of overweight	22.46	4.79E-06
Avg. Domestic private health expenditure per capita (current US$)	19.3	2.05E-05
Avg. Total alcohol consumption per capita	15.61	0.0001176
Avg. Population ages 0-14, total	5.68	0.01836

Number of Clusters: 2, **Number of Points**: 158, **Between-group Sum of Squares**: 26.356, **Within-group Sum of Squares**: 46.312, **Total Sum of Squares**: 72.668, **Cluster 1**: 48, **Cluster 2**: 110

Table 13. Health Predictors Over Time

Indicator Averages by Year	2005	2010	2015	2020	Delta
Health - Cluster 1					
Average of Life expectancy at birth, total (years)	57.776	60.398	63.188	64.272	11.24 %
Average of Lifetime risk of maternal death (%)	2.467	2.134	1.694	1.389	-43.70 %
Average of People practicing open defecation (% of population)	30.039	25.119	20.018	15.410	-48.70 %
Average of Prevalence of undernourishment (% of population)	22.589	18.762	17.904	17.487	-22.59 %
Average of Incidence of tuberculosis (per 100,000 people)	313.309	283.798	247.213	211.935	-32.36 %
Average of Number of maternal deaths	5023.505	4615.484	4148.356	3826.008	-23.84 %
Average of Prevalence of overweight (modeled estimate, % of children under 5)	5.096	4.217	4.131	4.312	-15.38 %
Average of Domestic private health expenditure per capita (current US$)	20.124	32.199	36.721	34.004	68.97 %

Average of Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)	3.680	3.979	4.107	3.951	7.35%
Average of Population ages 0-14, total	14943161.678	16078116.082	17217157.593	18314101.302	22.56 %
Health - Cluster 2					
Average of Life expectancy at birth, total (years)	73.653	75.028	76.177	76.162	3.41%
Average of Lifetime risk of maternal death (%)	0.149	0.122	0.104	0.082	-44.86 %
Average of People practicing open defecation (% of population)	2.237	1.518	0.914	1.082	-51.61 %
Average of Prevalence of undernourishment (% of population)	6.989	5.812	5.033	4.878	-30.20 %
Average of Incidence of tuberculosis (per 100,000 people)	57.780	52.811	48.150	35.509	-38.54 %
Average of Number of maternal deaths	251.881	221.000	182.283	158.854	-36.93 %
Average of Prevalence of overweight (modeled estimate, % of children under 5)	8.735	8.678	8.107	8.007	-8.34%
Average of Domestic private health expenditure per capita (current US$)	312.725	417.392	436.216	474.766	51.82 %
Average of Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)	6.778	6.661	6.587	6.142	-9.38%
Average of Population ages 0-14, total	6452706.302	6418491.733	6548088.810	6805112.455	5.46%

Discussion

Interpretation

The results were encouraging because they revealed statistically significant relationships and clearly addressed the main questions we set out to explore. The overall findings from the modeling analysis showed that some indicators had a stronger predictive effect on life expectancy than others and even quantified these effects. The overall findings from the clustering analyses underscore that countries can indeed be grouped into distinct categories based on environmental, socioeconomic, and health indicators. The results suggest that clusters with similar profiles tend to exhibit similar life expectancy trends. These insights not only reveal the inherent heterogeneity among nations but also highlight potential pathways and policy levers that might be targeted to improve public health outcomes. By identifying these significant world development indicators and tracking their evolution over time, we gain a more nuanced perspective on how different domains interact to shape population health, offering guidance for targeted interventions and policy development.

Limitations

Though some of the results do align with prevailing research or even common knowledge while other results require further research to understand why they have a strong predictive effect on life expectancy (e.g. “Fertilizer consumption” in the environmental model shown in Table 5 on page), some of the results unfortunately are rather counterintuitive. For example, the second-strongest indicator that has a positive relationship with life expectancy in the health model is “Number of maternal deaths,” as seen in Table 7 on page . It has a coefficient of .533, indicating a positive relationship such that for every one standard deviation increase in the

number of maternal deaths, life expectancy increases by .533 years. Such a result seems counterintuitive; one would expect a negative relationship instead.

This situation highlights a major limitation and challenge of applying analytics to World Bank datasets: analytics can uncover statistical associations that do not appear meaningful. As such, cautious interpretation is required when proffering the results of an analysis – while much of the results are actionable without needing deeper analysis, some of it needs to be investigated further before being deemed actionable.

The results of this study, then, has formed the basis of a starting point for future research: interesting indicators such as "Fertilizer consumption" could become the focus of sociological studies as related to health, while counterintuitive indicators such as "Number of maternal deaths" could become the focus of extended analytics projects that take a deeper dive using domain knowledge as well as advanced analytics techniques to untangle the complex web of hidden correlation and multicollinearity so that any counterintuitive result would have a logical and causal explanation for its inclusion in the final, actionable recommendations.

Conclusion

This study explored the relationship between life expectancy at birth and various world development indicators, utilizing a comprehensive dataset spanning 2004 to 2021. Through exploratory data analysis, ordinary least squares regression models, and k-means clustering, we identified the most influential socioeconomic, environmental, and health-related factors affecting life expectancy. Our findings highlight that health indicators, particularly those reflecting healthcare access and disease burden, have the strongest correlation with life expectancy, followed by socioeconomic and environmental factors. Furthermore, each domain,

socioeconomic, environmental, and health, was able to produce prediction models with meaningful explanatory power, as evidenced by the respective R-squared values of 0.806, 0.665, and 0.881. These results indicate that each set of indicators contributes significantly to predicting life expectancy at birth.

Additionally, the clustering analysis provided valuable insights into how countries with similar profiles in these indicators exhibit comparable life expectancy trends. These findings suggest that while investment in healthcare is critical, addressing broader socioeconomic and environmental factors is also important for improving population health outcomes. By utilizing the significant independent variables identified in each OLS model, policymakers now have key targets and levers that they can focus on to aid decision-making. These targeted interventions can help prioritize resources effectively and drive policies that address the most influential factors, ultimately contributing to healthier, more sustainable global development.

Works Cited

Group 1. *BANA 698 CULMINATING PROJECT (CSUN Fall 2025)*. 2025. Web. 28 April 2025. <https://github.com/CSUN-MS-BANA/culminating-project-group-1>.

MLA Handbook Plus. *1.5: Internal Headings and Subheadings*. 2022. Web. 28 April 2025. <https://doi.org/10.1632/GIJR2317>.

Office of Disease Preventation and Health Promotion. *Social Determinants of Health*. 2024. https://odphp.health.gov/healthypeople/priority-areas/social-determinants-health. 16 April 2025.

World Bank Group. *Metadata Glossary: SP.DYN.LE00.IN*. 2025. 16 April 2025. <https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SP.DYN.LE00.IN>.

—. *Quarterly Public Sector Debt (QPSD)*. 2025. Database. 27 January 2025. <https://databank.worldbank.org/source/quarterly-public-sector-debt>.

—. *Statistical Performance Indicators (SPI)*. 2025. Database. 27 January 2025. <https://databank.worldbank.org/source/statistical-performance-indicators-(spi)>.

—. *World Development Indicators (WDI)*. 2025. Database. 27 January 2025. <https://databank.worldbank.org/source/world-development-indicators#>.

World Health Organization. *Indicator Metadata Registry List*. 2022. 16 April 2025. <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/65>.

—. *World Health Statistics 2022*. 2022. 16 April 2025. <https://cdn.who.int/media/docs/default-source/gho-documents/world-health-statistic-reports/worldhealthstatistics_2022.pdf>.