Life Expectancy Prediction using Multiple Linear Regression

Sachith Nimesh, Hisal Perera, Shehani Thilakarathne and Lelumi Edirisinghe (Group 6)

BSc. (Hons) Financial Mathematics and Industrial Statistics Semester 4

University of Ruhuna

MIS2231: Case Study I

July 24, 2023

Introduction

Life expectancy is an important indicator to assess the population health of a country's entire population. Max Rose et al. (2013) state that life expectancy is the key metric to assess population health because it is broader compared to the narrow metric of infant and child mortality which only considers the mortality of a limited age group.

Life Expectancy

Definition: The World Health Organization(WHO) defines Life expectancy at birth as "The average number of years that a newborn could expect to live if he or she were to pass through life exposed to the sex- and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area." (Landry, n.d.) It is measured in years.

Life expectancy is a statistic used to measure the overall mortality level of the population across all age groups.

Method of Estimation: Life expectancy at birth is calculated using sex- and age-specific death rates obtained from life tables. The United Nations provides life expectancy at birth values correspond to mid-year estimates. They are consistent with the relevant United Nations fertility medium quinquennial population projections. Available mortality data from civil registration are used to build life tables. They are used to construct life tables after quality assessment and adjustments for completeness for registration. WHO provides a model life table based on a modified logit system obtained using about 1800 life tables. Those 1800 life tables have been obtained from reliable and essential sources. This model is used to plan the life tables needed to estimate life expectancy with a limited number of input parameters. Countries with annual life tables use a weighted regression model to project parameters. There more weight is given to recent data. Then the projected parameters are fed to the modified logit model using national data to predict complete life tables. In a situation where the age-specific mortality rates are insufficient, estimated under-5 mortality rates and estimated adult mortality rates or only the estimated under-5 mortality rates are used in life table derivation using a modified logit model with a global standard (average of all 1800 life tables) The output statistic is mainly a prediction.

Review of Literature on Life Expectancy

The review of literature on life expectancy as a proxy for a nation's health status is useful to investigate the factors that affect it. In this respect, this section is allocated to review the literature on determinants of a nation's life expectancy.

Hansen and Strulik (2015) found that that the cardiovascular revolution led to an increase in adult life expectancy by about 2 years, which caused higher education enrollment to increase by 7 percentage points across U.S. states.

Shin (2013) surveyed the impact of a pension system on life expectancy and the lifetime utility level. This study suggested that the pension system can make life expectancy longer or shorter and it is not always true that the pension system improves the lifetime utility level.

Hazan (2012) indicated a positive correlation between the percentage change in schooling and the change in life expectancy at birth during 1960-1990.

Balan and Jaba (2011) showed that the determinants with a positive impact on the life expectancy of the Roma population are wages, the number of beds in hospitals, the number of doctors, and the number of readers subscribed to libraries, while the determinants with a negative impact on life expectancy are the ratio Roma population and the ratio of the illiterate population for the year 2008.

Halicioglu (2010) investigated the factors of life expectancy in Turkey for the period 1965- 2005. In this study, the determinants of life expectancy in Turkey have been classified into selected economic, social, and environmental factors. According to the results of this study, the nutrition and food availability factors were the main positive factors for improving lifetime. But smoking was the main cause of mortality.

Bergh and Nilsson (2009) analyzed the relationship between three dimensions of globalization (economic, social, and political) and life expectancy using a panel of 92 countries over the period 1970-2005. They found a very robust positive effect from economic globalization on life expectancy, even when controlling for income, nutritional intake, literacy, number of physicians and several other factors.

Mariani et al. (2008) determined the relationship between life expectancy and environmental quality dynamics. The results showed environmental conditions affected life expectancy.

Yavari and Mehrnoosh (2006) analyzed the effects of socioeconomic factors on life expectancy using multiple regression analysis. This study showed that there is a positive, strong correlation between life expectancy as an independent variable and per capita income, health expenditures, literacy rate and daily calorie intake. Also, it revealed that there is a negative strong correlation between life expectancy and the number of people per doctor in African countries.

Leung and Wang (2003) investigated the relationship between health care, life expectancy and output using a modified neoclassical growth model. They showed income and economic development factors have positive impacts on lifetime.

Bernard et al. (2003) investigated the effects of saving behaviour on life expectancy. They indicated that a decrease in saving behaviour did not relate to an increase in individual life expectancy.

Castello and Domenech (2002) provided a theoretical model in which inequality affects per capita income when individuals decide to accumulate human capital depending on their life expectancy. According to the finding of this study, the distribution of education was dependent on the existence of multiple steady states.

Cervellati and Sunde (2002) investigated the relationship between human Capital Formation, life expectancy and the process of economic development, experienced by the Western world when passing from an environment of economic stagnation to sustained growth. The results indicated that human capital formation and life expectancy potentially reinforced each other due to advances in technological progress.

Summing up, the review of presented studies shows that the determinants of life expectancy can be divided into economic, social, environmental, and health-related factors.

Thus, in our case study, we try to develop a multiple linear regression model to predict the life expectancy of a country using some economic, social, and health-related factors from recent data (from 2000 to 2015) obtained from WHO for 193 countries.

Research Questions

1. What are the key social, economic, and health-related predictors useful to develop a multiple linear regression model to predict the life expectancy of a country?

2. What is the relative importance of each key predictor in predicting the life expectancy of a country?

3. Is the developed multiple regression model reliable in predicting the life expectancy of a country? What are the limitations?

2. Materials and Methods

Research Approach

The proposed overall research approach is the quantitative research approach.

Justification: The 3 research questions under the case study belong to convergent reasoning. By considering each research question, that claim can be justified as follows.

1. What are the key social, economic, and health-related predictors useful to develop a multiple linear regression model to predict the life expectancy of a country?

Out of all the social, economic, and health-related predictors in the data set, it is required to identify only the key predictors that are useful in developing a multiple linear regression model. Thus, by specifying only the useful predictors the possibilities are narrowed down. Therefore, this question belongs to convergent reasoning.

2. What is the relative importance of each key predictor in predicting the life expectancy of a country?

By considering the relative importance the specific contribution from each key predictor is determined. Therefore, this is convergent reasoning.

3. Is the developed multiple regression model reliable in predicting the life expectancy of a country? What are the limitations?

Examining model reliability and limitations converges towards conclusions about the utility and essential improvements of the model, thus an example of convergent reasoning.

Quantitative research is considered better for convergent reasoning due to the following reasons.

1. Objectivity: Quantitative research uses objective measurements, numerical data, and statistical analysis. The multiple linear regression model is one of the tools of the quantitative approach. It helps to specifically identify the key predictors and quantify their relative importance.

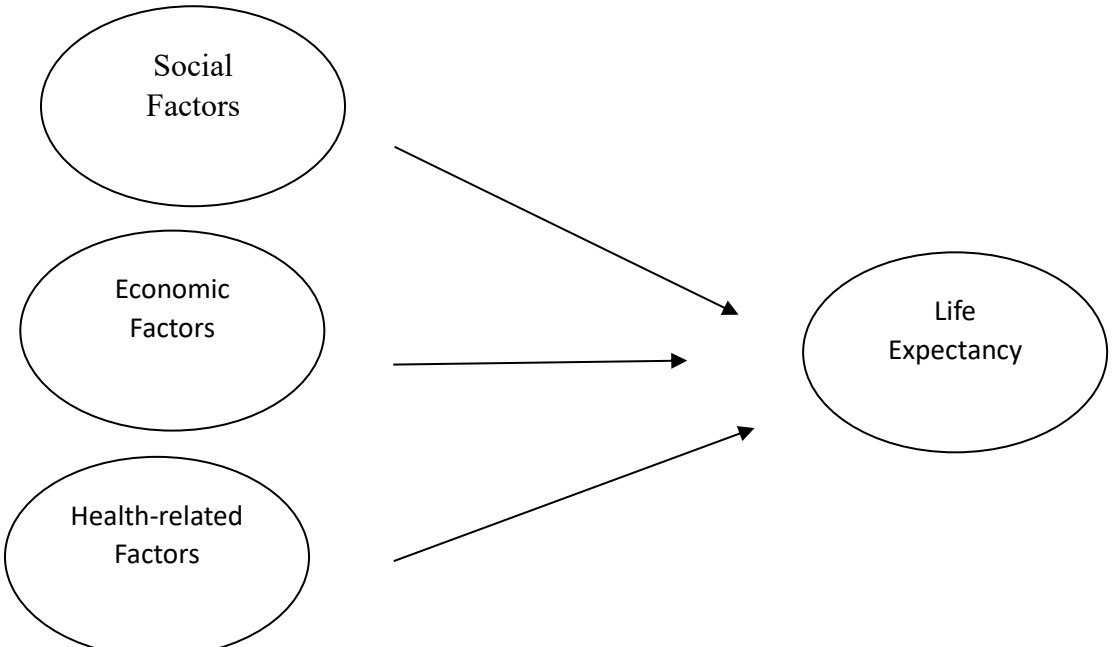
2. Data-Driven Decision-Making: Quantitative research depends on data-driven decisions which is important to obtain specific and measurable conclusions in convergent reasoning.

3. Statistical Inference: The quantitative approach facilitates statistical inference. The strength and significance of the obtained relationship can be examined.

4. Replicability: The quantitative approach is relatively transparent and replicable. Others can replicate the finding using the same or similar data set and the used statistical techniques such as multiple linear regression. This improves the credibility of the study.

Conceptual Model

Here is the general conceptual model representing broader categories of predictors under study. The relationships between life expectancy and the key predictors selected from each broader category will be explored during the case study.



Research Design

This is a quantitative case study which utilizes multiple linear regression.

Justification: Multiple linear regression enables quantifying the relationship of multiple social, economic, and health-related factors with life expectancy. It builds a mathematical model that can be used to predict the life expectancy of a country under certain assumptions. Using the same data set or a similar one, the multiple linear regression model can be replicated. It increases the credibility of the case study. Since multiple linear regression is a versatile statistical tool, we can utilize it for multiple tasks such as prediction, relationship analysis, variable selection, hypothesis testing, etc. which improves the scope of the study.

3. Data

Dataset

The dataset used for the case study has been obtained by the platform Kaggle. ((KUMARRAJARSHI, n.d.). The dataset comprises observations from 193 countries for the years 2000-2015. The data have been collected from the Global Health Observatory (GHO) data repository under the WHO and United Nation website.

Data Dictionary

Table 1

Data Dictionary: Variable Description, Variable Type, Measurement Units

Variable Name	Variable Description	Variable Type	Measurement Units
Country	Country observed	Categorical	No unit
Year	Year observed	Numerical	Calendar year

Status	Developed or developing status of the observed country	Categorical	No unit
Life Expectancy	Life Expectancy in age	Numerical	years
Adult Mortality	Adult mortality rates in both males and females	Numerical	No. of deaths per 1000 population aged 15-60 years.
Infant Deaths	No. of infant deaths per 1000 population	Numerical	No. of infant deaths per 1000 population
Alcohol	Recorded 15+ per capita alcohol consumption	Numerical	Liters of pure alcohol
Percentage Expenditure	Health expenditure as a percentage of GDP per capita	Numerical	Percentage (%)
Hepatitis B	Percentage of hepatitis B immunization coverage among 1-year-olds	Numerical	Percentage (%)
Measles	No. of reported measles cases per 1000 population	Numerical	No. of reported measles cases 1000 population
BMI	Average body mass index of the entire population of the observed country	Numerical	Kilograms per square meter (kg/m^2)

Under-Five Deaths	No. of under-five deaths per 1000 population	Numerical	No. of under-five deaths per 1000 population
Polio	Percentage of Polio (Pol3) immunization coverage among 1-year-olds	Numerical	Percentage (%)
Total Expenditure	General government health expenditure as a percentage of total expenditure	Numerical	Percentage (%)
Diphtheria	Percentage of Diphtheria tetanus toxoid and pertussis coverage among 1-year-olds	Numerical	Percentage (%)
HIV/AIDS	No. of deaths due to HIV/AIDS among 0–4-year-olds per 1000 live births	Numerical	No. of 0–4-year-olds HIV/AIDS deaths per 1000 live births
GDP	Gros Domestic Product per capita	Numerical	USD
Population	Population of the country	Numerical	No. of individuals
Thinness 10-19 years	Prevalence of thinness among age 10-19 years	Numerical	Percentage (%)

Thinness 5-9 years	Prevalence of thinness among age 5-9 years	Numerical	Percentage
Income Composition of Resources	Human Development Index in terms of income composition of resources	Numerical	No unit (index ranges from 0 to 1)
Schooling	Average no. of schooling years in the country	Numerical	years

Missing Data: Missing Data were mostly observed from relatively unknown countries (e.g.: Cook Islands, Saint Kitts and Nevis, Vanuatu), some small countries like Monaco and some countries with political dictatorships like North Korea and Sudan.

Missingness Map

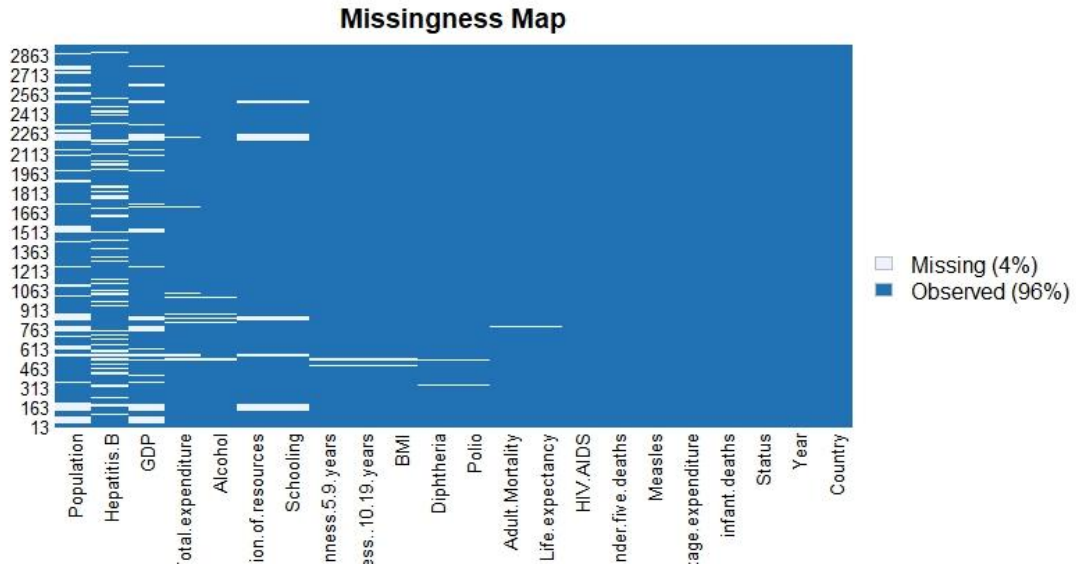


Figure 1

The missing values are mainly observed in variables such as Population, Hepatitis B, and GDP. It is not feasible to obtain reliable values for these missing values. The percentage of missing values is only about 4%. So, we decided to drop the missing observations assuming the removal of the unknown data will not affect the generalizability of the findings to most of the countries in the world. After removing the missing observations, the final dataset consists of 1649 observations of 22 variables.

Preparation For Analysis

Selecting Variables: The dataset includes 22 variables with 20 numerical and 2 categorical variables. Before the analysis, the variables should be filtered and mutated considering their observed properties in initial data preparation.

1. Remove the variables that do not provide additional information to predict life expectancy.

The categorical variable country has too many unique levels. (High cardinality problem). Each country's economic, social, and health-related data have been included as separate observations in other variables. Therefore, the country variable does not provide additional information beneficial to the study. Thus, remove it.The numerical variable Year is a time series data. Our study focuses on time-independent economic, social, and health-related predictors of life expectancy. Therefore, the Year variable does not provide a benefit to the study thus removing it.

2. Mutate the variables with wide ranges between the minimum and first quartile.

The ranges between the minimum and the first quartile of the variables Hepatitis-B, Polio and Diphtheria are too wide. To treat this, we decided to perform data discretization. We mutate the three numerical variables into three categorical variables with two categories covered under 90% and covered more than or equal to 90%. These two categories are designed in accordance with the threshold given by the Global Vaccine Action Plan (GVP) 2011-2020 that each country needs to reach above all or equal to 90% national coverage for all vaccines by 2020.

Metadata

Collaborator: KumarRajarshi

Sources: https://www.who.int/

https://www.worldbank.org/en/home

https://ourworldindata.org/

4. Exploratory data analysis

Exploratory Data Analysis (EDA) is a critical first stage in every data analysis project, including the life expectancy case study. Its goal is to obtain a deeper knowledge of the dataset, uncover patterns, linkages, and potential concerns, and lead us in developing hypotheses for future quantitative research. In the life expectancy case study, we will use EDA to better understand the dataset and highlight crucial insights.

Data Collection and Loading

Begin by obtaining a life expectancy dataset from a trustworthy source or database. The dataset should include factors relevant to life expectancy, such as nation, year, GDP, healthcare spending, and so on. Load the dataset into the R environment using appropriate libraries and review the first few rows to confirm proper loading. (Figure 2)

Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure	Hepatitis.B	Measles	BMI
Afghanistan	2015	Developing	65.0	263	62	0.01	71.279624	65	1154	19.1
Afghanistan	2014	Developing	59.9	271	64	0.01	73.523582	62	492	18.6
Afghanistan	2013	Developing	59.9	268	66	0.01	73.219243	64	430	18.1
Afghanistan	2012	Developing	59.5	272	69	0.01	78.184215	67	2787	17.6
Afghanistan	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	17.2
Afghanistan	2010	Developing	58.8	279	74	0.01	79.679367	66	1989	16.7

under.five.deaths	Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP	Population	thinness..1.19.years	thinness.5.9.years
83	6	8.16	65	0.1	584.25921	33736494	17.2	17.3
86	58	8.18	62	0.1	612.69651	327582	17.5	17.5
89	62	8.13	64	0.1	631.74498	31731688	17.7	17.7
93	67	8.52	67	0.1	669.95900	3696958	17.9	18.0
97	68	7.87	68	0.1	63.53723	2978599	18.2	18.2
102	66	9.20	66	0.1	553.32894	2883167	18.4	18.4

Income.composition.of.resources	Schooling
0.479	10.1
0.476	10.0
0.470	9.9
0.463	9.8

Figure 2

Data Summary

We have generated a numerical summary of the dataset to get an initial overview. This includes mean, median, standard deviation, minimum, and maximum values for relevant numerical variables (e.g., life expectancy, GDP, schooling, etc.).

Descriptive Statistics of variables

Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths
Length:1649	Min. :2000	Length:1649	Min. :44.0	Min. : 1.0	Min. : 0.00
Class :character	1st Qu.:2005	Class :character	1st Qu.:64.4	1st Qu.: 77.0	1st Qu.: 1.00
Mode :character	Median :2008	Mode :character	Median :71.7	Median :148.0	Median : 3.00
	Mean :2008		Mean :69.3	Mean :168.2	Mean : 32.55
	3rd Qu.:2011		3rd Qu.:75.0	3rd Qu.:227.0	3rd Qu.: 22.00
	Max. :2015		Max. :89.0	Max. :723.0	Max. :1600.00

Alcohol	percentage.expenditure	Hepatitis.B	Measles	BMI	under.five.deaths
Min. : 0.010	Min. : 0.00	Min. : 2.00	Min. : 0	Min. : 2.00	Min. : 0.00
1st Qu.: 0.810	1st Qu.: 37.44	1st Qu.:74.00	1st Qu.: 0	1st Qu.:19.50	1st Qu.: 1.00
Median : 3.790	Median : 145.10	Median :89.00	Median : 15	Median :43.70	Median : 4.00
Mean : 4.533	Mean : 698.97	Mean :79.22	Mean : 2224	Mean :38.13	Mean : 44.22
3rd Qu.: 7.340	3rd Qu.: 509.39	3rd Qu.:96.00	3rd Qu.: 373	3rd Qu.:55.80	3rd Qu.: 29.00
Max. :17.870	Max. :18961.35	Max. :99.00	Max. :131441	Max. :77.10	Max. :2100.00

Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP	Population
Min. : 3.00	Min. : 0.740	Min. : 2.00	Min. : 0.100	Min. : 1.68	Min. :3.400e+01
1st Qu.:81.00	1st Qu.: 4.410	1st Qu.:82.00	1st Qu.: 0.100	1st Qu.: 462.15	1st Qu.:1.919e+05
Median :93.00	Median : 5.840	Median :92.00	Median : 0.100	Median : 1592.57	Median :1.420e+06
Mean :83.56	Mean : 5.956	Mean :84.16	Mean : 1.984	Mean : 5566.03	Mean :1.465e+07
3rd Qu.:97.00	3rd Qu.: 7.470	3rd Qu.:97.00	3rd Qu.: 0.700	3rd Qu.: 4718.51	3rd Qu.:7.659e+06
Max. :99.00	Max. :14.390	Max. :99.00	Max. :50.600	Max. :119172.74	Max. :1.294e+09

thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources	Schooling
Min. : 0.100	Min. : 0.100	Min. :0.0000	Min. : 4.20
1st Qu.: 1.600	1st Qu.: 1.700	1st Qu.:0.5090	1st Qu.:10.30
Median : 3.000	Median : 3.200	Median :0.6730	Median :12.30
Mean : 4.851	Mean : 4.908	Mean :0.6316	Mean :12.12
3rd Qu.: 7.100	3rd Qu.: 7.100	3rd Qu.:0.7510	3rd Qu.:14.00
Max. :27.200	Max. :28.200	Max. :0.9360	Max. :20.70

Figure 3

Univariate Analysis

To display the distributions of crucial numerical data (such as life expectancy), we plotted histograms and violin plots. This will aid in determining whether they follow any pattern (normal, skewed, etc.). To display the distributions of crucial numerical data (such as life expectancy), we plotted histograms. This will aid in determining whether they follow any particular pattern (normal, skewed, etc.).

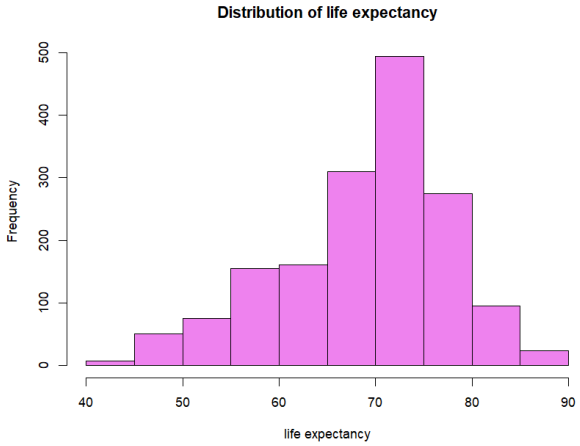


Figure 4

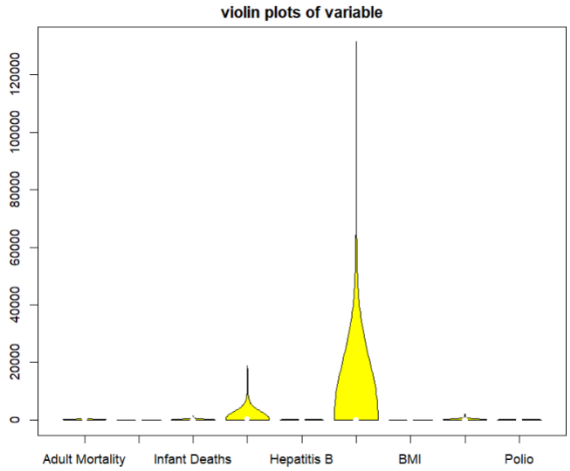


Figure 5

Histogram of independent variables

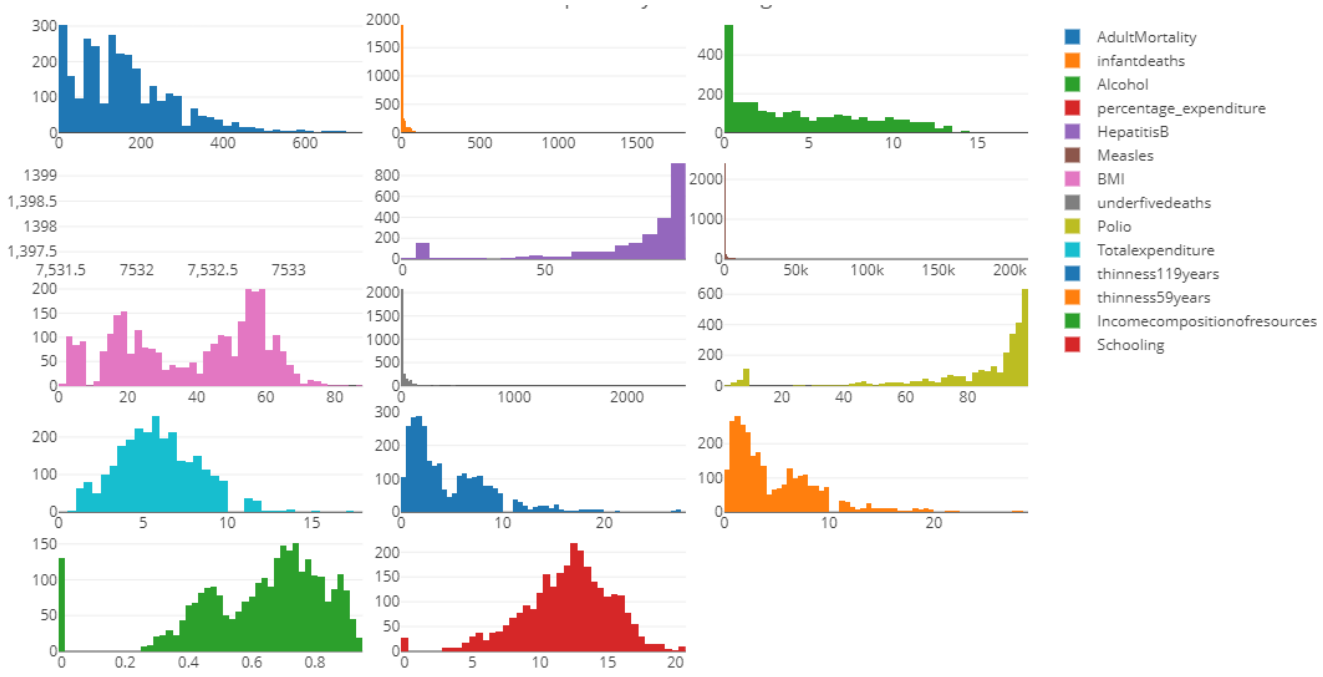


Figure 6

The histograms for life expectancy indicated a broadly normal distribution. We can see that there's many variations in this data set variables. This above dashboard shows skewness and structure of distribution.

Data Visualization

To uncover patterns of linkage between various numerical variables, we generated visualizations such as a correlation matrix and a heat map (Figure 4). Using interactive visualization tools to go deeper into data and unearth insights.

Correlation Matrix and Heat Map

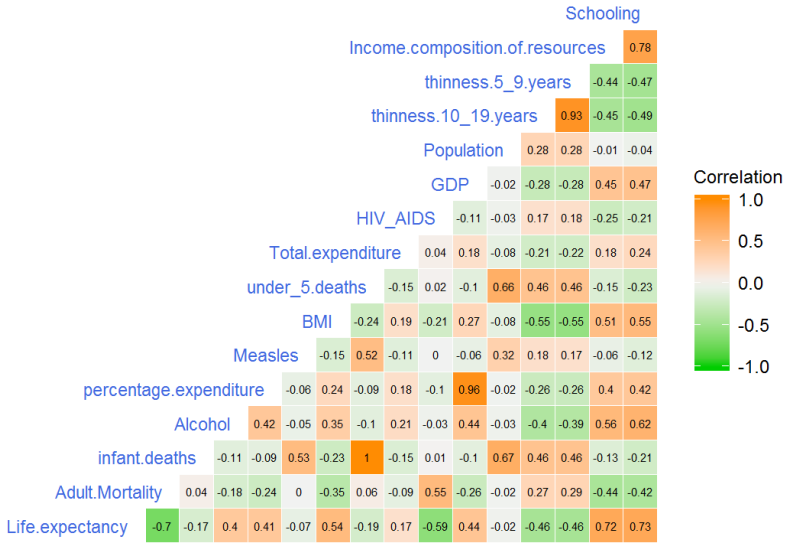


Figure 7

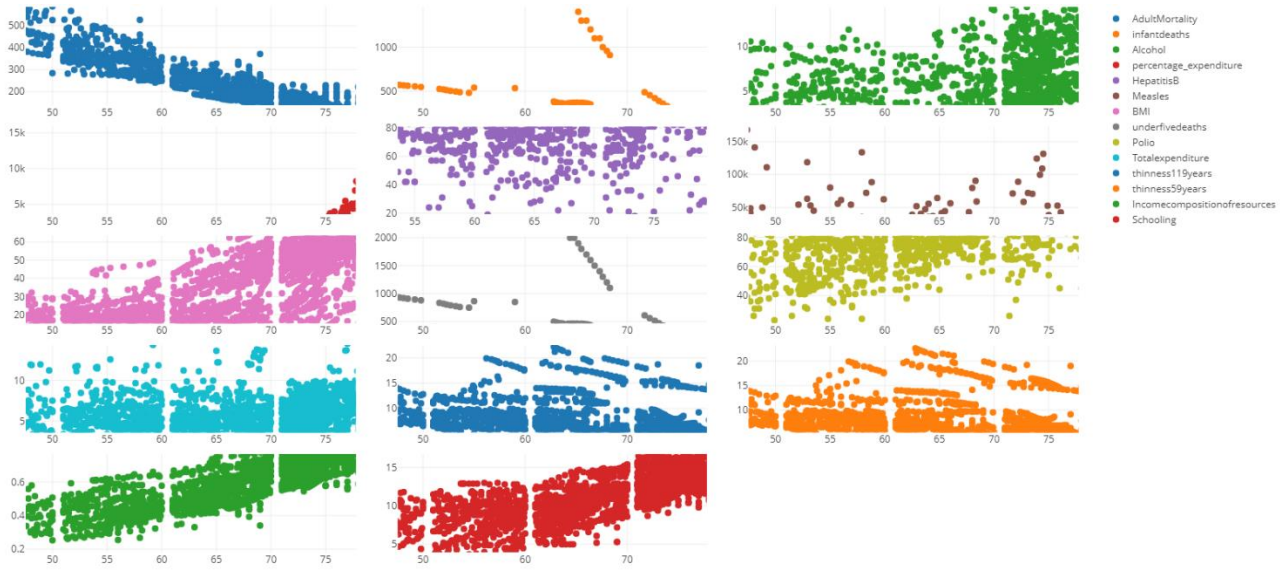


Figure 8 - Correlation between life expentancy and other variables respectively.

The life expectancy dataset exploratory data analysis offered useful insights that will guide our future quantitative research. We discovered intriguing trends and linkages, such as strong correlations between life expectancy and adult mortality, income comparison from resources and schooling while highlighting differences among nations and regions.

The analysis laid the groundwork for additional examination, allowing us to make educated judgments and obtain a better knowledge of the facts in this case study.

References

Roser , M., Ortiz-Ospina, E., & Ritchie, H. (2013). *Life Expectancy.* OurWorldInData.org.

Landry, M. M. (n.d.). *Life expectancy at birth*. Retrieved from World Health Organization: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3131

KUMARRAJARSHI. (n.d.). *Life Expectancy (WHO)*. Retrieved from kaggle: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Monsef, A. ., & Mehrjardi, A. S. . (2015). Determinants of Life Expectancy: A Panel Data Approach. *Asian Economic and Financial Review*, *5*(11), 1251–1257

Appendix

Extract of the dataset

The first 6 observations of life expectancy along with the considered social, economic, and health-related factors have been extracted from the data set.

Life Expectancy with Social Factors

Life expectancy	Alcohol	Schooling	Income composition of resources
65	0.01	10.1	0.479
59.9	0.01	10	0.476
59.9	0.01	9.9	0.47
59.5	0.01	9.8	0.463
59.2	0.01	9.5	0.454
58.8	0.01	9.2	0.448

The income composition of resources variable gives the Human Development Index in terms of income composition of resources considering inequalities and disparities of the population. Since it assesses fairness and inclusiveness of income distribution measuring social development in a sense, we decided to use it as a social factor instead of an economic factor.

Life Expectancy with Economic Factors

Life expectancy	percentage expenditure	Total expenditure	GDP
65	71.27962	8.16	584.2592

59.9	73.52358	8.18	612.6965
59.9	73.21924	8.13	631.745
59.5	78.18422	8.52	669.959
59.2	7.097109	7.87	63.53723
58.8	79.67937	9.2	553.3289

Life Expectancy with Health-related Factors

	thinness	thinness
Life	1-19	5-9
expectancy	years	years
65	17.2	17.3
59.9	17.5	17.5
59.9	17.7	17.7
59.5	17.9	18
59.2	18.2	18.2
58.8	18.4	18.4

Life expectancy	Adult Mortality	infant deaths	Hepatitis B	Measles	BMI	under-five deaths	Polio	Diphtheria	HIV/AIDS
65	263	62	65	1154	19.1	83	6	65	0.1
59.9	271	64	62	492	18.6	86	58	62	0.1
59.9	268	66	64	430	18.1	89	62	64	0.1
59.5	272	69	67	2787	17.6	93	67	67	0.1
59.2	275	71	68	3013	17.2	97	68	68	0.1
58.8	279	74	66	1989	16.7	102	66	66	0.1