Capstone Final Report

Name: H Karen Tam
Date: February 10, 2018

Executive Summary

Some studies suggested that people living in poverty have shorter life spans. From the equity standpoint, everyone should have the basic right to survive. Therefore, it is important to examine elements that are associated with life expectancy, and governments have responsibilities to close the gap of the inequality in terms of developing and implementing public policies.

This project evaluates how income, total and government health spending, improved drinking water source, food supply, and cancer death associate with life expectancy. The data of all variables are from the Gapminder.

Pearson Correlation Coefficient and Multiple linear regression were used to infer causal relationships between life expectancy and independent variables used in this project. Income, improved drinking water source, and food supply are related to life expectancy based on these two tests. Although cancer death is not significantly correlated with life expectancy, it is statistically significant to the regression model. On the other hand, total and government health spending are correlated with life expectancy but not statistically significant to the regression model.

Dataset

Data of seven variables from 1960 to 2016 are from the Gapminder, an independent Swedish foundation with no political, religious or economic affiliations. Each variable has its own dataset. A final dataset was created by combining datasets for the following variables:

Variable	Years	Number of Countries & Territories
Life Expectancy (Year)	1960 - 2016	209
GDP Per Capita (USD)	1960 - 2011	200
Total Health Spending Per Person (USD)	1995 - 2010	189
Government Health Spending Per Person (USD)	1995 - 2010	191
Improved Drinking Water Sources (Percentage of Population)	1990 - 2010	201
Food Supply (Calories)	1961 - 2007	176
Cancer Death	1995 - 2002	149

Limitations

There are many reasons for death. The dataset only contains very limited variables that are probably related to the life expectancy. This dataset cannot answer questions about relationships between life expectancy, diseases other than cancers and pollution.

Cleaning and Wrangling

- Reshape the data for plots and statistical analysis
    - I used gather() in tidyr package to gather year columns into key-value pairs. It was easier to make scatter plots and do statistical analysis by using the restructured dataset.
    - In each row, I added up numbers of different types of cancer and created a column to store the total number of cancer deaths.
- Remove missing value and join the data from different datasets
    - I removed missing values in each individual variable's dataset.
    - I used inner_join() in dplyr package to join all variable's datasets together into a large dataset. I also calculated total numbers of cancer death for each country in different years by adding numbers of various male and female cancer death.
- Examine outliers
    - I ran summary statistics and made subsets to identify outliers for each variable. Outliers were examined. These are reasonable with the countries' situations in specific years. For example, extremely short life expectancy is appeared due to wars and natural disasters in specific countries and in the specific time period. It was the fact that a few countries might have very high numbers of death of a particular type of cancer. Various cancer deaths and life expectancy are normally distributed. It is expected that income and health spending have more outliers because some countries are wealthier than other countries.

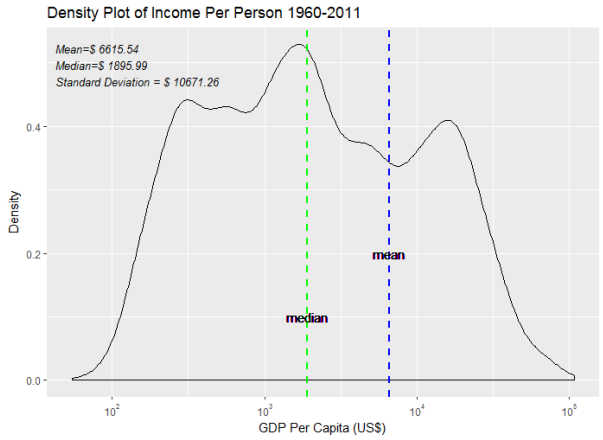
Preliminary Exploration and Initial Finding

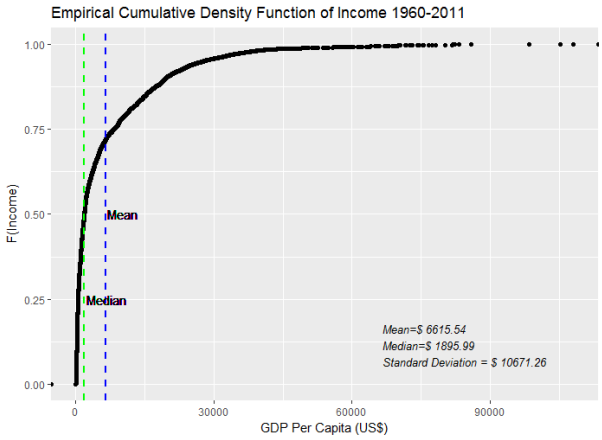
GDP Per Capita

Data from 200 countries and territories is used in plots below. GDP Per Capita by country is not normally distributed.

There are 1/3 of the countries with less than US$557 income per person. Nine of the ten countries with the lowest means of GDP Per Capita are from different parts of Africa. Nepal is the only one Asian country on the bottom ten list.

Outliers are countries with citizens who have very high income in average ($80,000). The top five countries with the highest means of GDP Per Capita are small countries and territories in Western Europe. The United Arab Emirates and Qatar in the Middle East, Japan in Asia, and United States in North America are among the top ten as well.





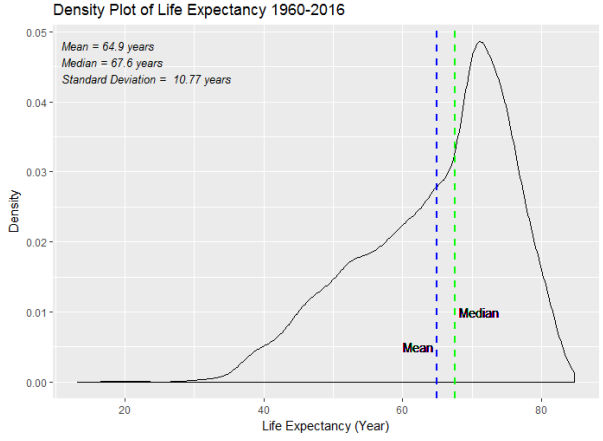
Observation: 7,988 without missing data

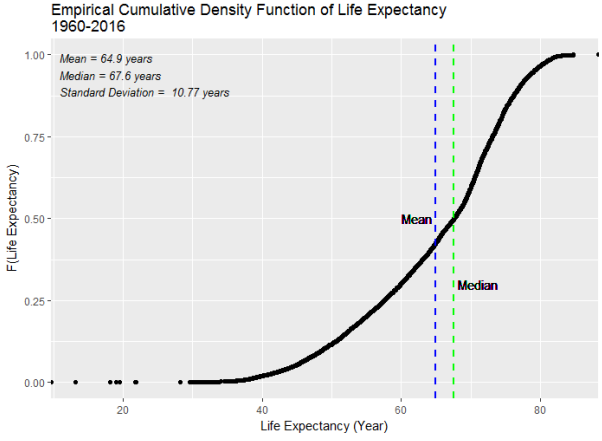
Source: Gapminder, Income per person (GDP/capita, PPP$ inflation-adjusted)

Life Expectancy

Data on life expectancy was from 209 countries and territories. The data is left-skewed, suggesting that the majority of countries have the long life expectancy.

Outliers are from countries with the life expectancy of fewer than 30 years of age during wars. They include Mali in 1960 and 1961. Since gaining independence from France in 1960, the West African state of Mali has been afflicted by several rebellions, insurrections, and coups. Other outliers were created by the genocide in Cambodia from 1975 to 1980, that resulted in the deaths of 25% of the country's population from starvation, overwork, and executions. Rwanda is one of these outliers as well due to the genocide in 1994. There were 500,000 to 1,000,000 deaths during that year.





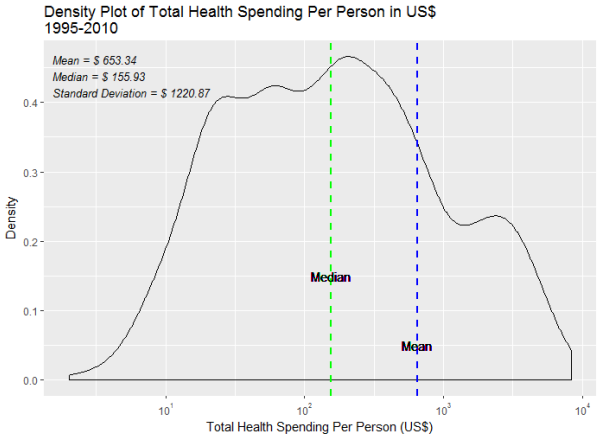
Observation: 11,630 without missing data

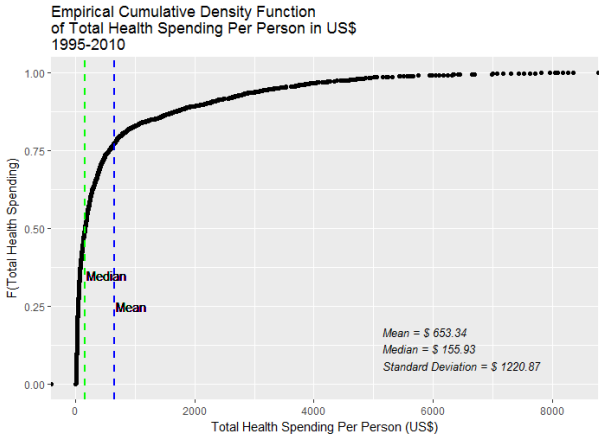
Source: Gapminder, Life Expectancy (Years)

Total Health Spending

Data on total health spending was from 189 countries and territories. It is not normally distributed and is right-skewed. There are 3/4 of countries and territories with the total health spending per person of US$564 or less.

Outliers are countries with total health spending more than US$8,000 per person per year. There are 6 records in this subset. The countries are Luxembourg, Norway, and United States from 2008 to 2010. These are rich countries, and their annual expenses are from $8,054 to $8,361.





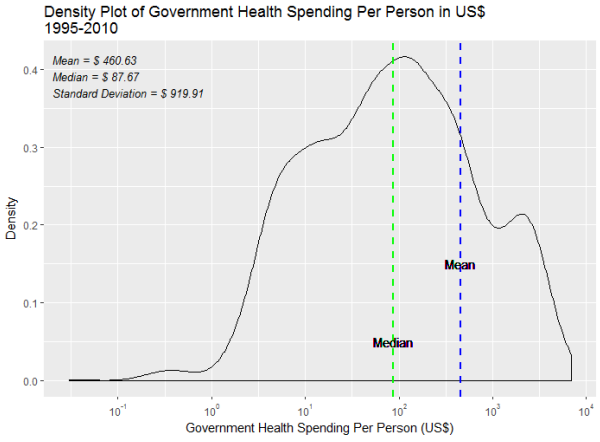
Observation: 3,008 without missing data.

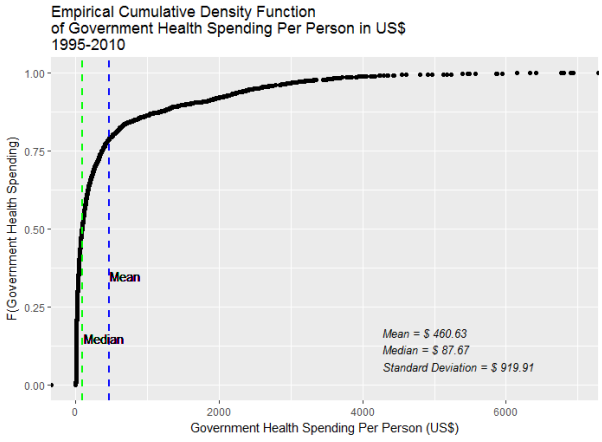
Source: Gapminder, Total health spending per person (USD)

Government Health Spending

Data on government health spending was from 191 countries and territories. Similar to the total health spending, it is not normally distributed and is right-skewed. There are about 3/4 of the countries with the government health spending per person of US$460 or less.

Outliers are Luxembourg and Norway from 2007 to 2010 with more than US$6,000 government health spending per person per year.





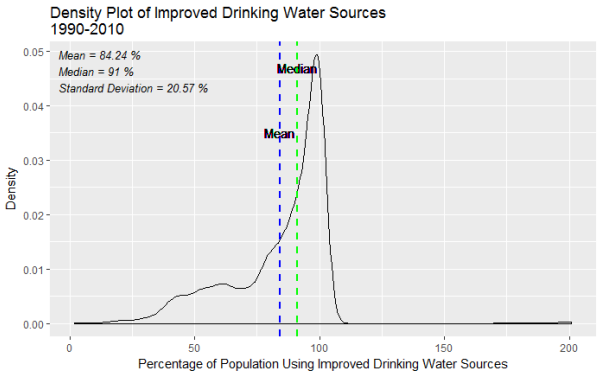
Observation: 3,039 without missing data

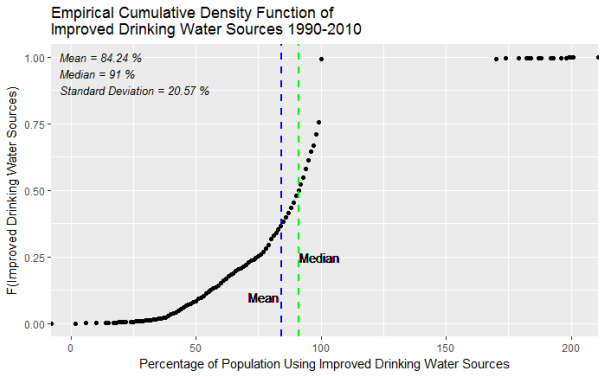
Source: Gapminder, Government health spending per person (USD)

Improved Drinking Water Sources

Data on improved drinking water sources was from 201 countries and territories. There are 106 countries and territories with more than 90% of the population using improved drinking water sources.

Outliers are from Afghanistan in different years in the 1990s with less than 10% of the population using improved drinking water sources. In 2007, Afghanistan has improved the accessibility to clean drinking water to 50% of its population.





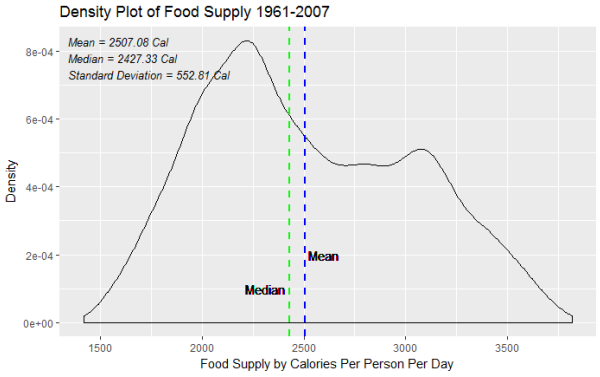
Observation: 4,045 without missing data.

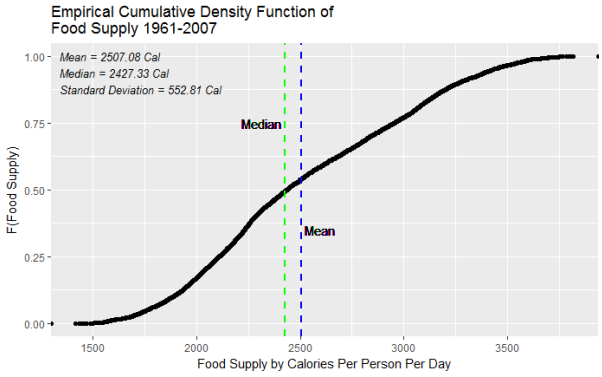
Source: Gapminder, Improved water source-overall access (%)

Food Supply

Data on average daily intake was from 176 countries and territories. According to the National Health Service of the United Kingdom, the average male adult needs approximately 2,500 calories per day to keep his weight constant, while the average adult female needs 2,000 calories per day. In this regard, the mean and median are close to recommended daily calories intakes.

There are only 6 countries with the average daily intake of fewer than 1,500 calories, but those are from the 1960s to 1990s. Countries include China, Burkina Faso, the Maldives in the 1960s; Ethiopia and Djibouti in the 1970s, and Eritrea in the 1990s. Outliers are countries with 3,750 calories per day per person or more, including the United States from 2002 to 2006, Hungary in 1987 and 1989, and Austria in 1999, 2000, 2001, 2006, and 2007.





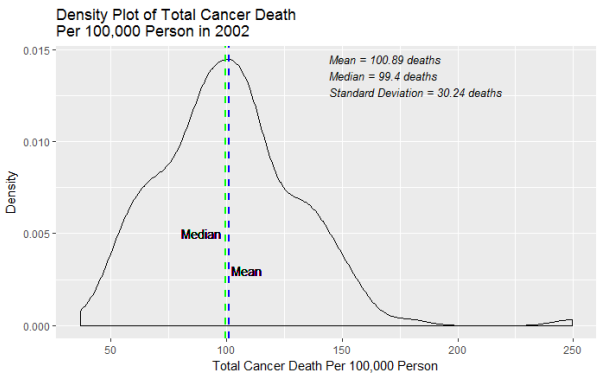
Observation: 7,500 without missing data

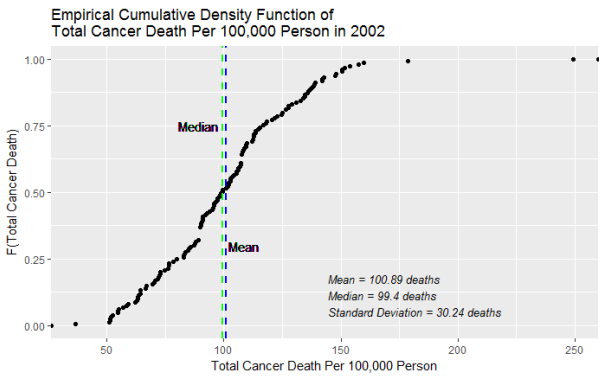
Source: Gapminder, Food supply (kilocalories/person & day)

Cancer Death

Data on various cancer death were from 149 countries and territories in 2002. The most updated available data is 2002's. The types of cancer include breast, cervical, liver, stomach, lung, and prostate. The data is almost normally distributed but is still right-stewed.

The outlier with 249.4 cancer deaths per 100,000 person per year in 2002 is Mongolia due to its extremely high rate of liver cancer (140.6 deaths per 100,000 person). Mongolia has the world's highest rate of liver cancer mortality. High rates of hepatitis C and B infection along with widespread alcohol use have left Mongolia with a burden of liver cancer that it is ill-equipped to handle.



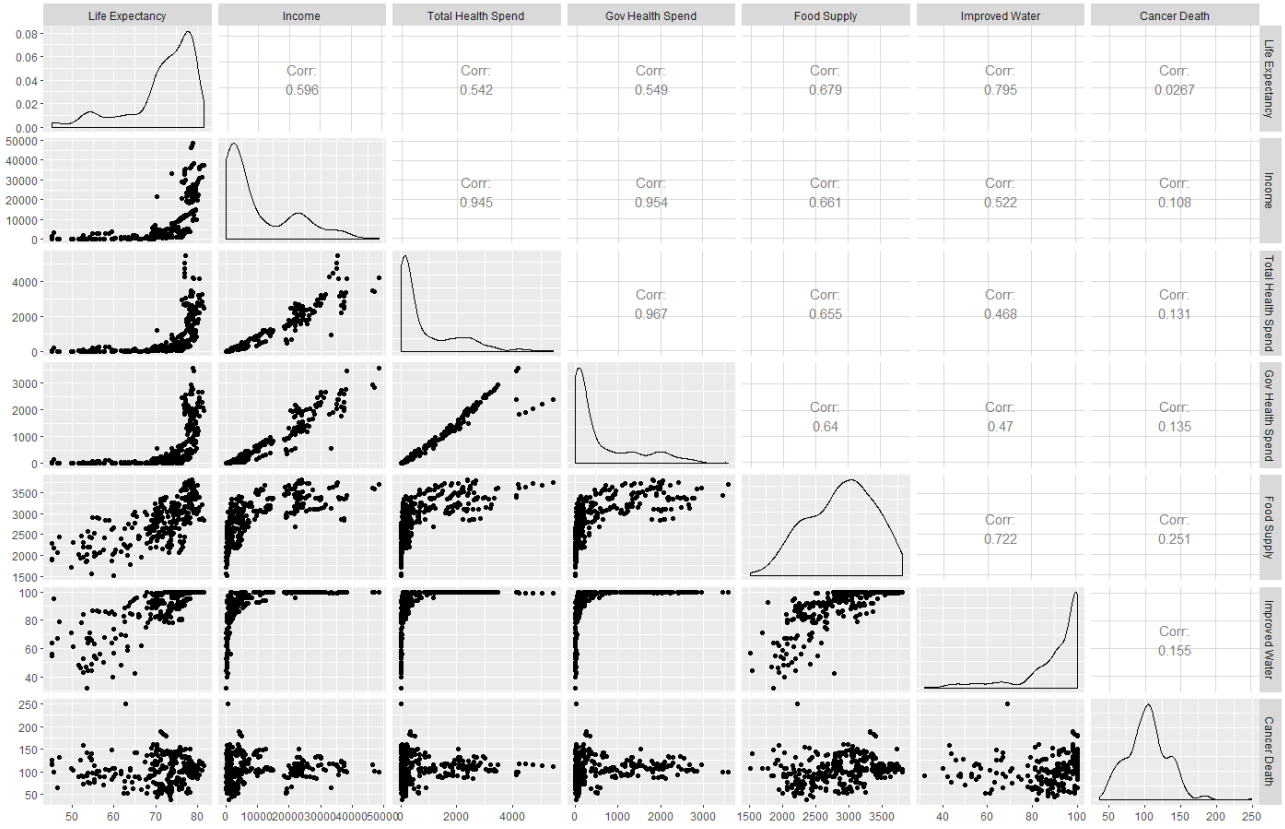


Observation: 149 without missing data

Source: Gapminder, Cancer Deaths Per 100,000 Persons (Search Keyword: Cancer)

Pearson Correlation Coefficient

The following is the pairs plot showing the pairwise comparisons and correlations of all 7 variables.



Observation: 346 without missing data.

From Year 1998 to 2002

Source: Gapminder

Summary of Correlations

Independent Variable	r (from Year 1998 to 2002)	Significantly Correlated with Life Expectancy	Strength of the Relationship	Direction of the Relationship
Income	0.60	Yes	Moderate	Positive
Total Health Spending	0.54	Yes	Moderate	Positive
Government Health Spending	0.55	Yes	Moderate	Positive
Improved Drinking Water Source	0.80	Yes	Strong	Positive
Food Supply	0.68	Yes	Strong	Positive
Cancer Death	-0.03	No	Weak	Negative

Scatterplot and explanation for each pairwise comparison between life expectancy and 6 independent variables is available here.

Multiple Linear Regression

Multiple Linear Regression, k-fold Cross-Validation, and three variable selection methods are used to identify the best model to estimate life expectancy.

Dataset

The dataset used in this test contains the most updated available data from 1998 to 2002 from the Gapminder. A total of 346 observations (countries/territories) are in the dataset. There is no missing data.

Steps To Create the Best Regresson Model

Variable Selection

- Best Subsets Regression
- Stepwise Forward Regression
- Random Forest

No.	Variable Seletion Method	Strong Independent Variables	Result
1	Best Subsets Regression	1. Income 2. Improved Drinking Water Source 3. Food Supply 4. Cancer Death	Adjusted R-Squared: 0.69
2	Stepwise Forward Regression	1. Income 2. Improved Drinking Water Source	Adjusted R-Squared: 0.69

		3. Food Supply 4. Cancer Death	
3	Random Forest	1. Income 2. Improved Drinking Water Source 3. Total Health Spending 4. Government Health Spending	1. Mean of Squared Residuals of out-of-bag predictions versus targets: 13.8 2. Percentage of Variable Explained: 78.5 3. Pseudo R-Squared: 0.78

The adjusted R-squared gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable.

Mean of Squared Residuals is a measure of the discrepancy between the data and an estimation model.

Pseudo R-Squared ranges from 0 to 1 with higher values indicating better model fit.

k-fold Cross-Validation

10-fold Cross-Validation is performed for 3 models with different combinations of variables. The combinations of variables used are based on recommendations from the above variable selection methods.

No.	Model Containing the Following Independent Variables	Root Mean Squared Error (RMSE)
1	1. Income 2. Improved Drinking Water Source 3. Food Supply 4. Cancer Death 5. Total Health Spending 6. Government Health Spending	4.46
2	1. Income 2. Improved Drinking Water Source 3. Food Supply 4. Cancer Death	4.42
3	1. Income 2. Improved Drinking Water Source 3. Total Health Spending 4. Government Health Spending	4.51

RMSE is used to measure differences between the values estimated by a model and the values actually observed.

Although Model 2 has the lowest RMSE of 4.42, all three models have similar RMSE. It indicates that they have similar accuracy of estimating life expectancy,

Compare Models with Different Independent Variables

No.	Combination of Independent Variables	Method(s) that Support(s) This Combination	Adjusted R-squared	Sum of Squared Errors (SSE)	Residual Standard Error
1	1. Income 2. Improved Drinking Water Source 3. Food Supply 4. Cancer Death 5. Total Health Spending	N/A	0.69	6762	4.47 years

	6. Government Health Spending				
2	1. Income 2. Improved Drinking Water Source 3. Food Supply 4. Cancer Death	1. Best Subsets Regression 2. Stepwise Forward Regression	0.69	6779	4.46 years
3	1. Income 2. Improved Drinking Water Source 3. Total Health Spending 4. Government Health Spending	Random Forest	0.67	7162	4.58 years

The adjusted R-squared gives the percentage of variation explained by only those Independent variables that in reality affect the dependent variable.

Sum of Squared Errors (SSE) is the sum of the squared differences between each observation and its group's mean.

Residual Standard Error is a measure of how close the fit is to the points.

Selected by Best Subsets Regression and Stepwise Forward Regression, Model 2 has the lowest Residual Standard Error of 4.46 years. Its SSE of 6779 is very close to Model 1's SSE, the lowest one in the table. The Adjusted R-Squared of Model 1 and 2 are almost the same. Model 2 is selected to be the estimator of this project.

Interpretation of the Selected Model

The selected model includes 4 significant independent variables: income, food supply, improved drinking water source, and cancer death

Adjusted R-Squared: 0.69
69% of the variation in life expectancy can be explained by this model.

P-Value:
Income, food supply, improved drinking water source, and cancer death are statistically significant because their p-values are less than 0.05. Statistically significant means that the result is likely not due to random chance.

Residual Standard Error:
Residual Standard Error of 4.46 shows that there are about 4.46 years between the observed life expectancy and estimated life expectancy.

Impacts on Life Expectancy:

Variable	Amount Needed to Increase 1 Year in Life Expectancy
Income	An Increase of $7,003 in GDP Per Capita
Improved Drinking Water Source	An increase in 3% of country's population with an access to clean drinking water sources
Food Supply	An increase in 453 calories intake per person per day (Note: If people faces food shortage due to extreme poverty or wars)
Cancer Death	A decrease in 29 cases of cancer death per 100,000 person per year

Conclusion

The following table shows the summarized results of the Pearson Correlation and the Multiple Linear Regression:

Independent Variable	Correlation	Regression
Income	Correlated	Significant
Improved Drinking Water Source	Correlated	Significant
Food Supply	Correlated	Significant
Cancer Death	Not Correlated	Significant
Total Health Spending	Correlated	Not Significant
Government Health Spending	Correlated	Not Significant

Recommendations

From the equity standpoint, country and local governments, public health researchers, heathcare facilities, and community nonprofits are critical to close the inequality gap in life expectancy between the rich and the poor. Recommendations for these stakeholders are as follows:

Stakeholders	Recommended Actions
All Stakeholders	Identify life expectancy related problems in underprivileged communities
Governments	Develop policies and implement them effectively
Community Organizations and Health Facilities	Provide services to reduce life expectancy related problems in underprivileged communities

Appendices

A. R Code of the Project
B. Scatterplot and Correlation Coefficient for Each Pairwise Comparison between Life Expectancy and Independent Variables
C. R Code and Output of the Pearson Correlation Coefficient
D. R Code and Output of the Multiple Linear Regression
E. Video Presentation
F. slide Deck
G. PDF file of the Final Report