Data Visualization of Life Expectancy and Social Economic

Team: Ruijie Li (B00894018) + Junyuan Wu (B00836490) + Zezhou Lu (B00887213)

2025-04-08

Introduction

In this data report, we will explore and analyze a widely used dataset in statistics and data science—the "Life Expectancy & Socio-Economic (World Bank)" dataset. This dataset collects key health and economic indicators from approximately 70 countries around the world, including citizens' life expectancy, health conditions, and government expenditures in public health, education, and other areas.

Research questions are:

1. **First Question**: Which factors best distinguish countries with better health conditions from those with poorer health conditions? What structural differences exist between these countries in terms of health investment and basic sanitation services?
2. **Second Question**: Is there an association between `Life Expectancy` and `Health Expenditure` , `sanitation` condition? How do `Health Expenditure` and `Sanitation` condition influence the `Life Expectancy` in different countries?
3. **Third Question**: Is there have relationship between Health Expenditure and Disability-Adjusted Life Years (DALYs)?

Data

1. Overview

The Life expectancy & Socio-Economic (world bank) dataset is a well-known dataset in statistics and data science. It contains life expectancy and health conditions of citizen as well as government expenditure in society in about 70 countries around the world from 2001 to 2019.

2. Source & Collection

- **Source**: The data was extracted from World Bank Open Data.
- **Original Purpose**: Compare life expectancy across different countries and analyze how improvement in economic growth (GDP per capita) influence performance.
- **Collection Method**: The data was likely compiled from population census, governments.

3. Variables & Descriptions

The dataset contains **1296 observations** (countries in 19 years) and **15 variables** (11 numeric, 4 categorical).

Variable	Description	Units	Type
`Country.Name`	country name	-	Character
`Country.Code`	abbreviation of country name	-	Character
`Region`	region of the world country is located in	-	Character
`IncomeGroup`	country's income class	-	Character
`Year`	2001-2019 (both included)	year	Numeric
`Life.Expectancy.World.Bank`	Life expectancy at birth	years old	Numeric
`Prevelance.of.Undernourishment`	percentage of the population whose habitual food consumption is insufficient to provide the dietary energy levels	%	Numeric
`CO2`	Carbon dioxide emissions	kiloton	Numeric
`Health.Expenditure..`	Level of current health expenditure expressed as a percentage of GDP	%	Numeric
`Education.Expenditure..`	General government expenditure on education (percentage of GDP)	%	Numeric
`Unemployment`	Unemployment refers to the % share of the labor force	%	Numeric

Variable	Description	Units	Type
Sanitation	People using safely managed sanitation services (% of the population)	%	Numeric
Injuries	Disability-Adjusted Life Years (DALYs) due to Injuries (the number of years of healthy life lost due to illness or injury)	years	Numeric
Communicable	Disability-Adjusted Life Years (DALYs) due to Communicable diseases	years	Numeric
NonCommunicable	Disability-Adjusted Life Years (DALYs) due to Non-Communicable diseases	years	Numeric

4. Key Characteristics**

· Data Volume: The dataset contains 1,296 observations and 15 variables.
· Time Span: The data covers the years 2001 to 2019.
· Geographical Coverage: Observations from approximately 70 countries.
· Variable Composition: The dataset includes key economic and health indicators such as life expectancy, CO2 emissions, health expenditure, education expenditure, unemployment rate, among others.
· Variable Distribution: Some variables, like life expectancy and CO2 emissions, show significant differences between countries, indicating considerable variability in the data.

5. Why Was This Data Collected?

· Original Purpose: To compare life expectancy across different countries and analyze the impact of economic growth and environmental conditions (e.g., CO2 emissions) on public health.
· Application Areas: This dataset is commonly used in statistical analysis, regression modeling, machine learning, and public policy research.
· Significance: By comparing the data, it helps to understand the intrinsic relationships between economic development, environmental factors, and public health. The dataset also serves as a foundation for policy-making and resource allocation, as well as for teaching statistical modeling and data science.

6. Limitations & Considerations

```
· Data Integrity: Some variables may have missing, for example:
some country missing in the period 2001-2019.
· Time Span: The data only covers 2001 to 2019, which may not reflect
trends from earlier or more recent periods.
· Measurement Error: Differences in standards and methods used by
various countries for statistical indicators (e.g., health expenditure,
CO2 emissions) can lead to some measurement errors when comparing data.
· Lack of Metadata: The dataset lacks detailed information on collection
conditions, such as testing environments and statistical methods,
which may affect the interpretation of the results.
· Sample Representativeness: The dataset covers a limited number of
countries, which may not fully represent the global situation.
```

Exploratory data analysis

```
df <- read.csv("life expectancy.csv", stringsAsFactors = FALSE)
df$Corruption <- NULL # Remove the column with the most null values
df <- na.omit(df)
write.csv(df, "life_cleaned.csv", row.names = FALSE)
data <- read.csv("life_cleaned.csv")
head(data)
```

```
##   Country.Name Country.Code                     Region         IncomeGroup Yea
r
## 1      Albania          ALB     Europe & Central Asia Upper middle income 200
1
## 2    Argentina          ARG Latin America & Caribbean Upper middle income 200
1
## 3      Armenia          ARM     Europe & Central Asia Upper middle income 200
1
## 4      Austria          AUT     Europe & Central Asia         High income 200
1
## 5   Azerbaijan          AZE     Europe & Central Asia Upper middle income 200
1
## 6   Bangladesh          BGD                South Asia Lower middle income 200
1
##   Life.Expectancy.World.Bank Prevelance.of.Undernourishment    CO2
## 1                   74.28800                             4.9   3230
## 2                   73.75500                             3.0 125260
## 3                   71.80000                            26.1   3600
## 4                   78.57561                             2.5  67910
## 5                   67.05400                            17.0  26400
## 6                   65.95600                            15.9  25780
##   Health.Expenditure.. Education.Expenditure.. Unemployment Sanitation
## 1             7.139524                 3.45870       18.575   40.52090
## 2             8.371798                 4.83374       17.320   48.05400
## 3             4.645627                 2.46944       10.912   46.35190
## 4             9.269429                 5.57548        4.010   99.67940
## 5             2.755907                 3.50342       10.910   19.59670
## 6             2.063751                 2.17193        3.617   18.73701
##    Injuries Communicable NonCommunicable
## 1  117081.7    140894.78        532324.8
## 2 1397676.1   1507068.98       8070909.5
## 3  103371.8    122238.13        767916.2
## 4  240208.9     77701.17       2101883.6
## 5  235307.7    904186.52       1816141.2
## 6 5106399.3  29348014.24      20529108.2
```

Summary table

We create a summary table that displays the main statistics (range, median, mean, quartiles, etc.) of the key quantitative variables in the dataset: Life Expectancy (World Bank), CO2, Health Expenditure (%) and Communicable.

```
data %>%
  select(Life.Expectancy.World.Bank, CO2, Health.Expenditure.., Communicable)
%>%
  summary() %>%
  kable(format = "html", digits = 2, caption = "Summary Statistics") %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

Summary Statistics

Life.Expectancy.World.Bank	CO2	Health.Expenditure..	Communicable
Min. :40.37	Min. : 60	Min. : 1.821	Min. : 3133
1st Qu.:70.03	1st Qu.: 7357	1st Qu.: 4.856	1st Qu.: 83103
Median :74.75	Median : 37185	Median : 6.783	Median : 342867
Mean :72.82	Mean : 238629	Mean : 6.796	Mean : 4748258
3rd Qu.:79.26	3rd Qu.: 95440	3rd Qu.: 8.564	3rd Qu.: 2425909
Max. :84.21	Max. :10707220	Max. :20.413	Max. :254933810

Interpretation: This shows the mean is 72.82 years old, median is 74.75 years old, and ranges from 40.37 to 84.21 for Life.Expectancy. And also shows the ranges for CO2 is from 60 kiloton to 10707220 kiloton. The mean of Health.Expenditure is 6.796 % of total GDP. And the median of the number of years of healthy life lost due to communicable illness is 342867 years.

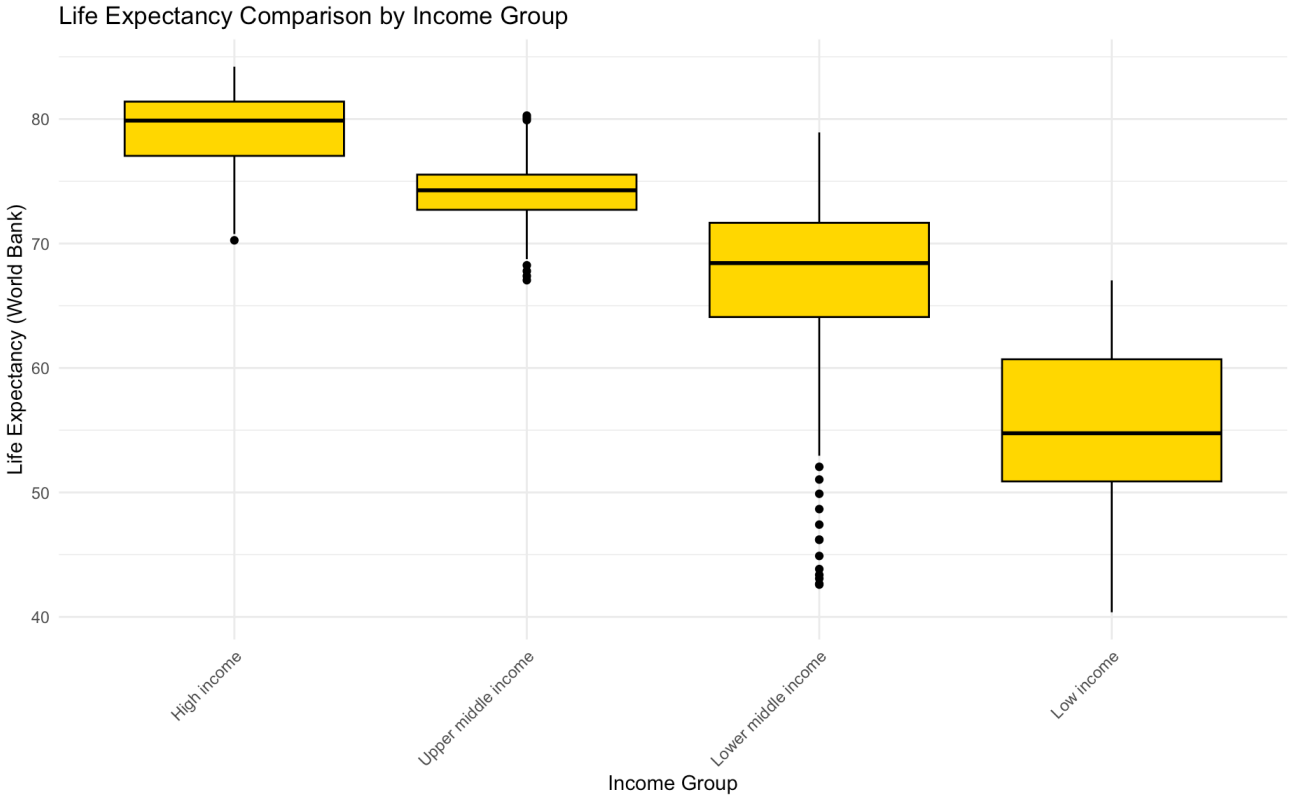
---

1. Life Expectancy Comparison by Income Group

Firstly, we use a box plot to compare life expectancy across different income groups. This visualization helps us assess whether economic status might be linked to variations in public health outcomes.

[Data-visualization-1: boxplot, Author=Junyuan Wu]

```
data$IncomeGroup <- factor(data$IncomeGroup,
                           levels = c("High income", "Upper middle income", "Low
er middle income", "Low income"))

ggplot(data, aes(x = IncomeGroup, y = `Life.Expectancy.World.Bank`)) +
  geom_boxplot(fill = "gold", color = "black") +
  theme_minimal() +
  labs(title = "Life Expectancy Comparison by Income Group",
       x = "Income Group",
       y = "Life Expectancy (World Bank)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Insight: The box plot reveals the distribution of life expectancy within each income group. Differences in the medians and spread across income groups may suggest that economic factors influence overall health outcomes.

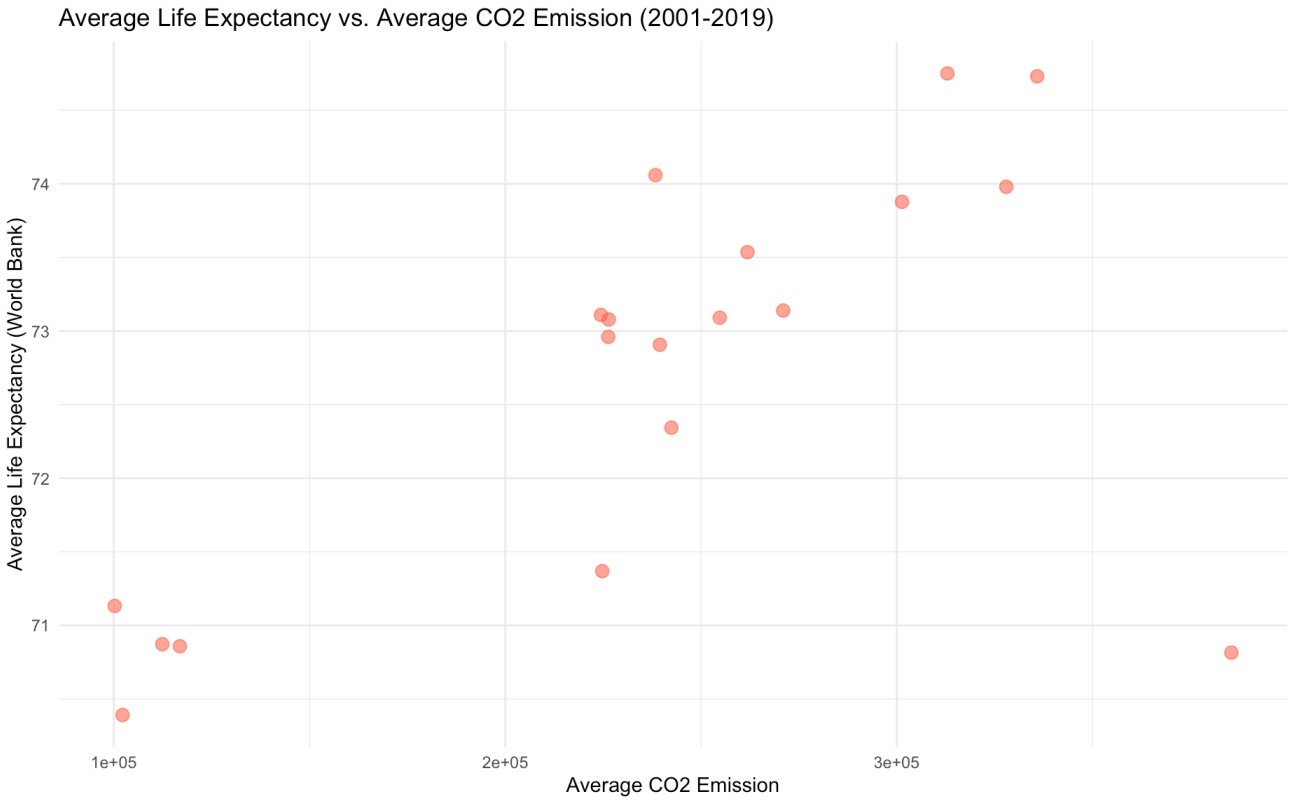
---

2. Average Life Expectancy vs. Average CO2 Emission (2001-2019)

Next, we explore the relationship between environmental factors and health by aggregating data from 2001 to 2019. We calculate the annual average life expectancy and average CO2 emission, and then display their relationship using a scatter plot.

[Data-visualization-2: scatter plot, Author=Junyuan Wu]

```
data_filtered <- data %>%
  filter(Year >= 2001 & Year <= 2019)

data_agg <- data_filtered %>%
  group_by(Year) %>%
  summarise(avg_life_expectancy = mean(`Life.Expectancy.World.Bank`, na.rm = TRU
E),
            avg_CO2 = mean(CO2, na.rm = TRUE))

ggplot(data_agg, aes(x = avg_CO2, y = avg_life_expectancy)) +
  geom_point(alpha = 0.6, color = "tomato", size = 3) +
  theme_minimal() +
  labs(title = "Average Life Expectancy vs. Average CO2 Emission (2001-2019)",
       x = "Average CO2 Emission",
       y = "Average Life Expectancy (World Bank)")
```



Insight: The scatter plot illustrates the relationship between average CO2 emissions and average life expectancy over time. This visualization may reveal whether higher environmental pollution is associated with lower life expectancy.

---

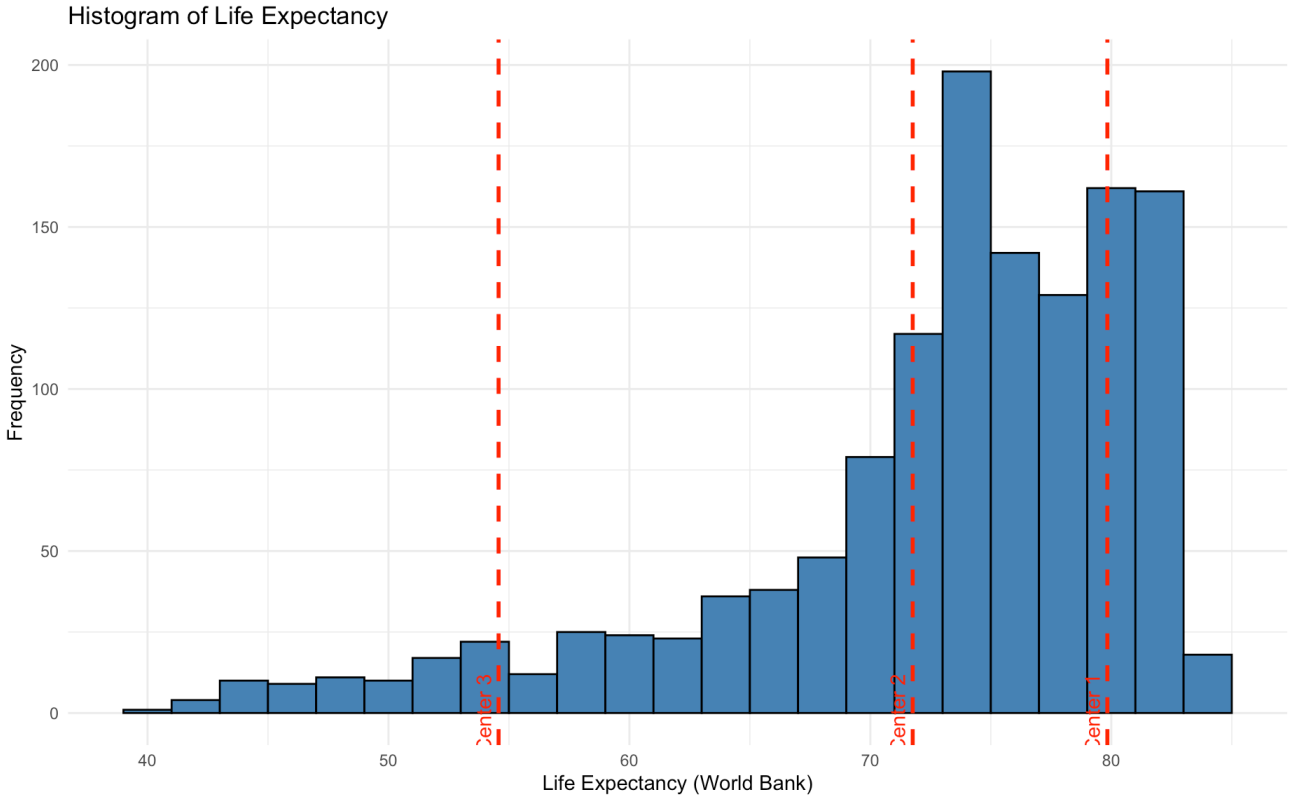
3. Histogram of Life Expectancy

Let us visualize the distribution of the Life Expectancy (World Bank) variable using a histogram. This plot helps us understand the overall distribution, identify any skewness, and spot potential outliers.

[Data-visualization-3: histogram, Author=Junyuan Wu]

```
data <- data %>%
  filter(!is.na(`Life.Expectancy.World.Bank`))

set.seed(123)
km <- kmeans(data$`Life.Expectancy.World.Bank`, centers = 3)
centers <- km$centers

ggplot(data, aes(x = `Life.Expectancy.World.Bank`)) +
  geom_histogram(binwidth = 2, fill = "steelblue", color = "black") +
  geom_vline(xintercept = centers, color = "red", linetype = "dashed", size = 1)
+
  geom_text(data = data.frame(x = centers),
            aes(x = x, y = 0, label = paste("Center", 1:length(centers))),
            angle = 90, vjust = -0.5, color = "red") +
  theme_minimal() +
  labs(title = "Histogram of Life Expectancy",
       x = "Life Expectancy (World Bank)",
       y = "Frequency")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Insight: The histogram shows the frequency distribution of life expectancy values. The red dashed lines represent the cluster centers obtained from a k-means clustering, which can help us understand how

the data groups naturally.

Analysis

Research Questions

1. **First Question**: Which factors best distinguish countries with better health conditions from those with poorer health conditions? What structural differences exist between these countries in terms of health investment and basic sanitation services?
2. **Second Question**: Is there an association between `Life Expectancy` and `Health Expenditure`, `sanitation` condition? How do `Health Expenditure` and `Sanitation` condition influence the `Life Expectancy` in different countries?
3. **Third Question**: Is there have relationship between Health Expenditure and Disability-Adjusted Life Years (DALYs)?

---

1. PCA for Multivariate Patterns： Average Health and Sanitation Conditions for Different Countries (2001–2019)

To explore multivariate relationships among key public health indicators, we performed Principal Component Analysis (PCA) on average national values of:Life Expectancy, Health Expenditure, Sanitation, Prevalence of Undernourishment. PCA helps reduce these four dimensions into two principal components (PC1 and PC2), capturing the majority of the variance across countries.

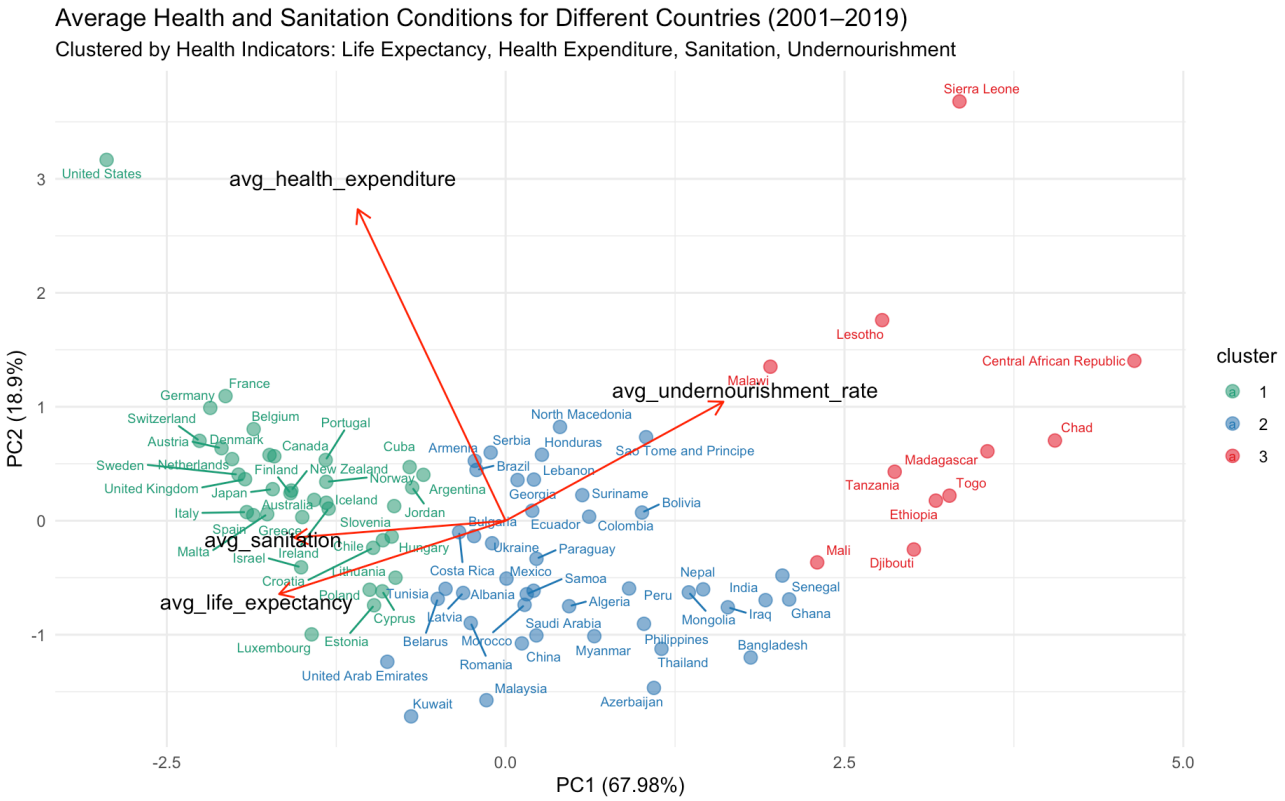
We also applied k-means clustering (k=3) in the PC1–PC2 space to group countries with similar health profiles.

PCA Visualization (Biplot): The biplot shows countries as points, colored by cluster. Red arrows represent raw variables and their contribution to each principal component (direction and magnitude)

[Data-visualization-4: biplot of PCA, Author=Zezhou Lu]

```
df_avg <- df %>%
  group_by(Country.Name) %>%
  summarise(
    avg_life_expectancy = mean(Life.Expectancy.World.Bank),
    avg_health_expenditure = mean(Health.Expenditure..),
    avg_sanitation = mean(Sanitation),
    avg_undernourishment_rate = mean(Prevelance.of.Undernourishment),
  )

# PCA analsis except the non-quantitative variable country name
pca <- prcomp(df_avg[, -1], center = TRUE, scale. = TRUE)

# PC1 and PC2, add country name variable into new pca data set.
pca_data <- as.data.frame(pca$x[, 1:2])
pca_data$Country <- df_avg$Country.Name

# Kmeans divided countries into into 3 clusters by pc1 and pc2
set.seed(123)
kcluster <- kmeans(pca_data[, c("PC1", "PC2")], centers = 3)
# Add cluster column into pca data set
pca_data$cluster <- as.factor(kcluster$cluster)

explained_var <- summary(pca)$importance[2, 1:2]
pc1_var <- round(explained_var[1] * 100, 2)
pc2_var <- round(explained_var[2] * 100, 2)

loadings <- as.data.frame(pca$rotation[, 1:2])
loadings$varname <- rownames(loadings)
loadings$PC1 <- loadings$PC1 * 3
loadings$PC2 <- loadings$PC2 * 3

ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster, label = Country)) +
  geom_point(size = 3, alpha = 0.6) +
  geom_text_repel(size = 2.5, max.overlaps = Inf) +
  geom_segment(data = loadings,
               aes(x = 0, y = 0, xend = PC1, yend = PC2),
               arrow = arrow(length = unit(0.3, "cm")),
               color = "red",
               inherit.aes = FALSE) +
  geom_text(data = loadings,
            aes(x = PC1 * 1.1, y = PC2 * 1.1, label = varname),
            color = "black", size = 4,
            inherit.aes = FALSE) +
  labs(
    title = "Average Health and Sanitation Conditions for Different Countries (2
001–2019)",
    subtitle = "Clustered by Health Indicators: Life Expectancy, Health Expendit
ure, Sanitation, Undernourishment",
    x = paste0("PC1 (", pc1_var, "%)"),
    y = paste0("PC2 (", pc2_var, "%)")
  ) +
```

```
  scale_color_manual(values = c("1" = "#1b9e77", "2" = "#1F78B4", "3" = "#E31A1
C")) +
  theme_minimal()
```



Key Insights:

- PC1 (67.98% of variance) reflects a “general health and development” axis.
    - Positively associated with Life Expectancy, Sanitation, and Health Expenditure
    - Negatively associated with Undernourishment
    - Countries on the right are generally healthier, with better sanitation and higher expenditure.
- PC2 (18.9% of variance) captures secondary contrasts, for example, some countries may have decent life expectancy despite lower health spending, or vice versa.
- Clusters reveal:

Cluster 1 (e.g., Norway, Japan, Germany): High health expenditure, high sanitation, low undernourishment – developed, high-health-status countries

Cluster 2 (e.g., India, Indonesia): Mid-range health metrics – transitional economies

Cluster 3 (e.g., Chad, Niger): Low sanitation, low expenditure, high undernourishment – developing nations with health challenges

Conclusion: This PCA approach provides a data-driven classification of countries based on multidimensional health indicators. It reveals how closely health expenditure, life expectancy, sanitation, and undernourishment are intertwined — and which countries diverge from expected trends.

This model supports international development planning, allowing stakeholders to identify peer countries, target aid, or benchmark progress.

2. Regression Analysis: Average Life Expectancy vs. Average Health Expenditure (2001–2019)

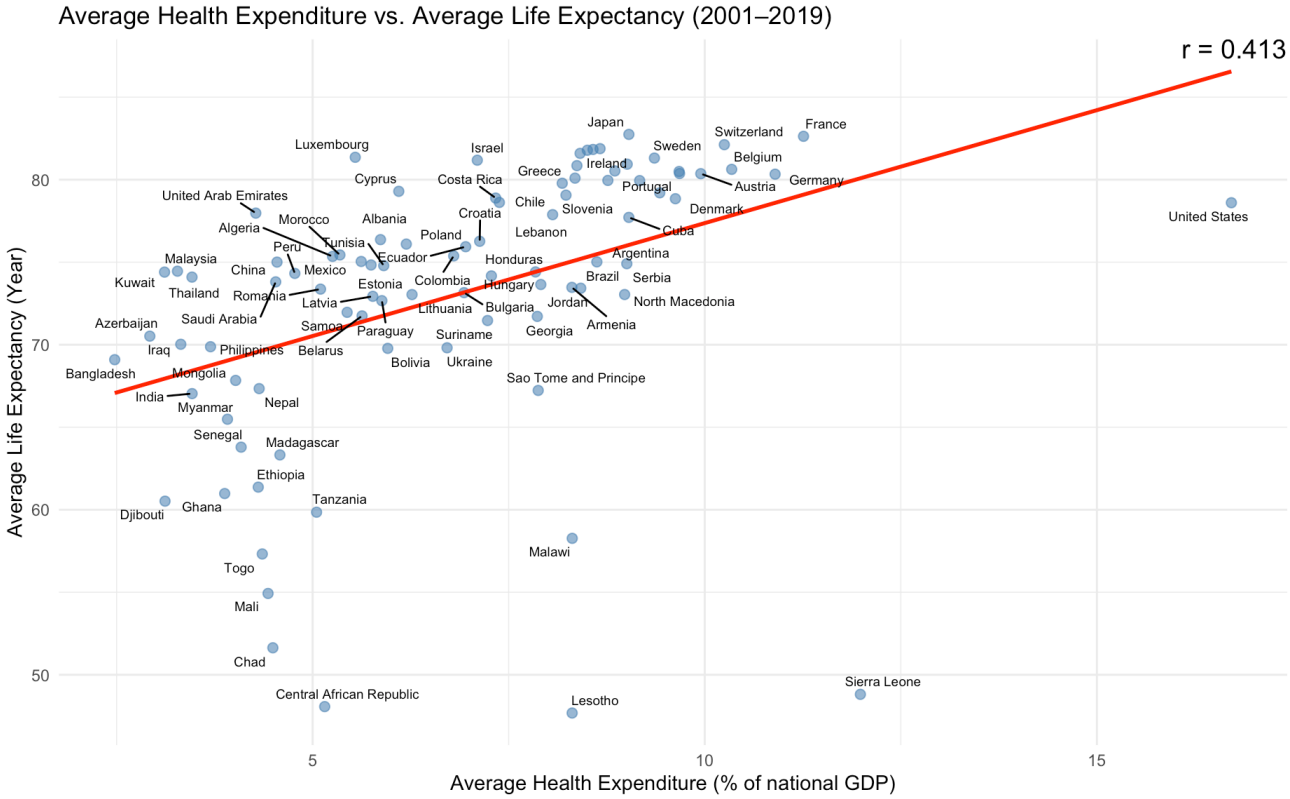
We use the regression visualization to identify the relationship between `Average Health Expenditure` and `Average Life Expectancy` across most of countries from 2001 to 2019. It fits **linear regression line** and also shows the correlation coefficient on the visualization.

[Data-visualization-5: Regression plot Life.Expectancy vs Health.Expenditure. Author=Zezhou Lu]

```
## Group by countries, and calculate each countries' the total average life expe
ctancy and average health       expenditure from 2001 to 2019
total_avg_health_expenditure <- data %>%
  group_by(Country.Name) %>%
  summarise(
    avg_life_expectancy = mean(Life.Expectancy.World.Bank),
    avg_health_expenditure = mean(Health.Expenditure..)
  )

R1 <- cor(total_avg_health_expenditure$avg_health_expenditure,
         total_avg_health_expenditure$avg_life_expectancy)

ggplot(total_avg_health_expenditure, aes(x = avg_health_expenditure, y = avg_lif
e_expectancy)) +
  geom_point(size = 2.2, alpha = 0.6, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_text_repel(aes(label = Country.Name), size = 2.5) +
  labs(
    title = "Average Health Expenditure vs. Average Life Expectancy (2001–201
9)",
    x = "Average Health Expenditure (% of national GDP)",
    y = "Average Life Expectancy (Year)") +
  annotate("text",
    x = Inf, y = Inf,
    hjust = 1, vjust = 1,
    label = paste0("r = ", round(R1, 3)),
    size = 5, color = "black") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Key Insights:

- Strong positive correlation: The Pearson correlation coefficient r ≈ 0.413 indicates a strong positive relationship
- Countries that invest more in health tend to have longer life expectancy.
- Linear trend: The red regression line shows a clear upward slope, meaning that increased health spending is generally associated with increased life expectancy.
- Some high-spending countries, for example the U.S., do not achieve proportionally higher life expectancy.
- Most of countries perform relatively well in the regression plot: more health expenditure corresponds to higher life expectancy.

Conclusion: Average health expenditure is a strong predictor of life expectancy, but the relationship is not perfectly proportional.

---

3. Regression Analysis: Average Life Expectancy vs. Average Health Expenditure (2001–2019)

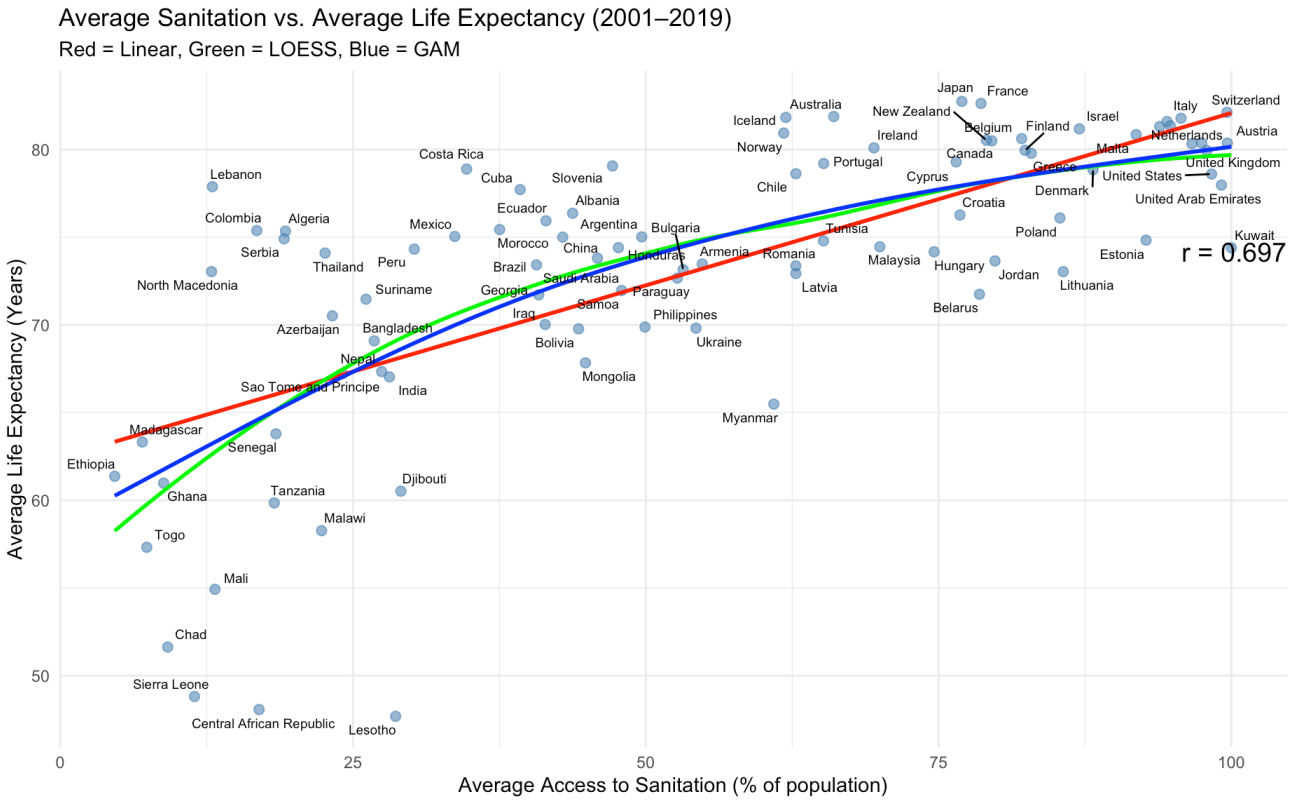
We use the regression visualization to identify the relationship between `Average Sanitation` and `Average Life Expectancy` across most of countries from 2001 to 2019. It fits **linear regression line, GAM and LOESS** and also shows thecorrelation coefficient on the visualization.

[Data-visualization-6: Regression plot with LOESS smooth model of sanitation vs life expectancy. Author=Zezhou Lu]

```
avg_sanitation <- data %>%
  group_by(Country.Name) %>%
  summarise(
    avg_life_expectancy = mean(Life.Expectancy.World.Bank),
    avg_sanitation = mean(Sanitation)
  )

R2 <- cor(avg_sanitation$avg_sanitation, avg_sanitation$avg_life_expectancy)

ggplot(avg_sanitation, aes(x = avg_sanitation, y = avg_life_expectancy)) +
  geom_point(size = 2.2, alpha = 0.6, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_smooth(method = "loess", se = FALSE, color = "green") +
  geom_smooth(method = "gam", formula = y ~ s(x), se = FALSE, color = "blue") +
  geom_text_repel(aes(label = Country.Name), size = 2.5) +
  labs(
    title = "Average Sanitation vs. Average Life Expectancy (2001–2019)",
    subtitle = "Red = Linear, Green = LOESS, Blue = GAM",
    x = "Average Access to Sanitation (% of population)",
    y = "Average Life Expectancy (Years)"
  ) +
  annotate("text",
    x = Inf, y = Inf,
    hjust = 1, vjust = 10,
    label = paste0("r = ", round(R2, 3)),
    size = 5, color = "black")  +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: ggrepel: 4 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Key Insights:

- Strong positive correlation: The correlation coefficient r ≈ 0.697 suggests a strong, positive relationship
- Countries with better sanitation access tend to have longer life expectancy.
- Linear fit (red): Shows a general upward trend — life expectancy increases as sanitation improves.
- LOESS (green) & GAM (blue): Capture nonlinear effects:
    - Gains in life expectancy are steepest when sanitation improves from low to moderate levels (from 30% to 70% coverage).
    - The curve flattens at high levels (above 85%), suggesting diminishing returns in highly developed countries.

Conclusion: Sanitation is a strong and nonlinear predictor of life expectancy — improving access in low-coverage regions can yield significant health benefits.

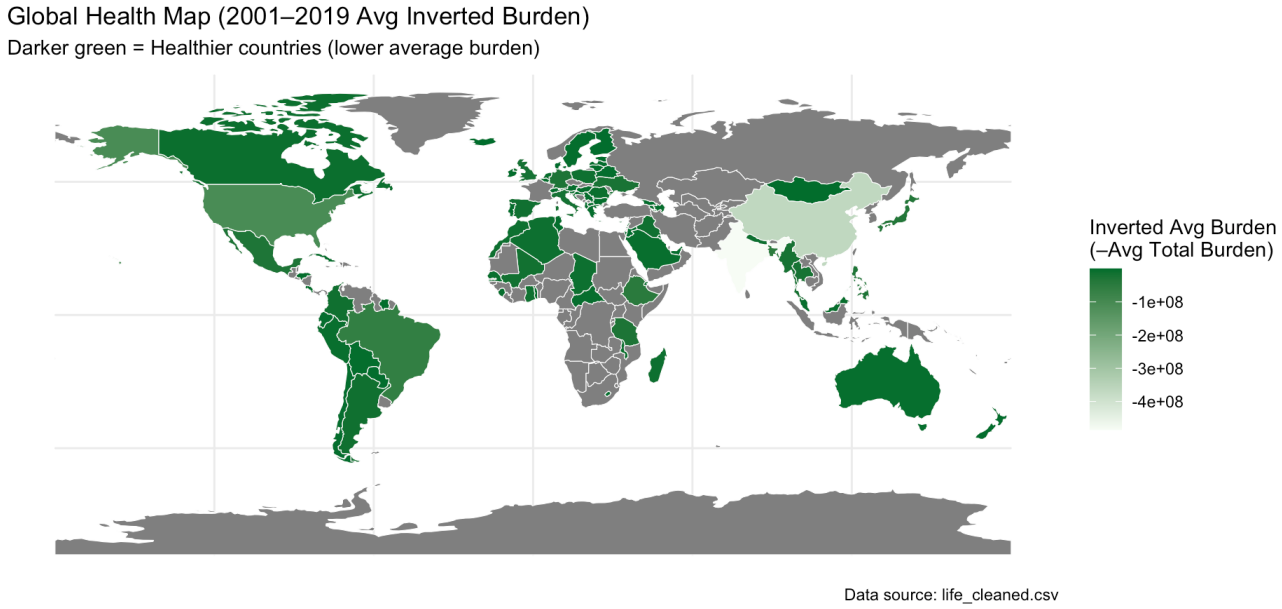
---

4. Final Model-Map Model: Global Health Map of Nagative Avg Inverted Burden (2001-2019)

To explore and visualize global health disparities, we constructed a spatial model mapping the average health burden (2001–2019) for each country. Here, a negative burden score (i.e., the inverted total health burden) is used so that darker green indicates better health outcomes.

[Data-visualization-7: Map Model of Injuries, Communicable, NonCommunicable VS country name. Author=Ruijie Li]

```
# calculate Health Burden Calculation
df <- data %>%
  mutate(
    Total_Burden = Injuries + Communicable + NonCommunicable) %>%
  group_by(Country.Code, Year, Country.Name)

# calculate Average Burden (2001–2019) of each country
avg_burden <- df %>%
  filter(Year >= 2001, Year <= 2019) %>%
  group_by(Country.Code, Country.Name) %>%
  summarise(Average_Burden = mean(Total_Burden, na.rm = TRUE), .groups = "drop")

# Load world map
world <- ne_countries(returnclass = "sf")

# Merge map data with value
map_data <- left_join(world, avg_burden, by = c("iso_a3" = "Country.Code"))

# change to negative Total_Burden
map_data$Neg_Burden <- -map_data$Average_Burden

# plot Green inverse map (the greener the healthier)
ggplot(map_data) +
  geom_sf(aes(fill = Neg_Burden), color = "white", size = 0.1) +
  scale_fill_gradient(
    low = "#f7fcf5",
    high = "#006d2c",
    name = "Inverted Avg Burden\n(–Avg Total Burden)"
  ) +
  labs(
    title = "Global Health Map (2001–2019 Avg Inverted Burden)",
    subtitle = "Darker green = Healthier countries (lower average burden)",
    caption = "Data source: life_cleaned.csv"
  ) +
  theme_minimal()
```



Key Insights:

- Countries in Sub-Saharan Africa and parts of South Asia show lighter shades, indicating higher average health burdens.
- Developed regions such as Western Europe, North America, and Australia tend to have darker green colors, reflecting stronger healthcare systems and lower burden levels.

Conclusion: This model supports health-related decision making by showing spatial inequalities in disease and injury burden. It also provides a foundation for comparing health investment needs across countries

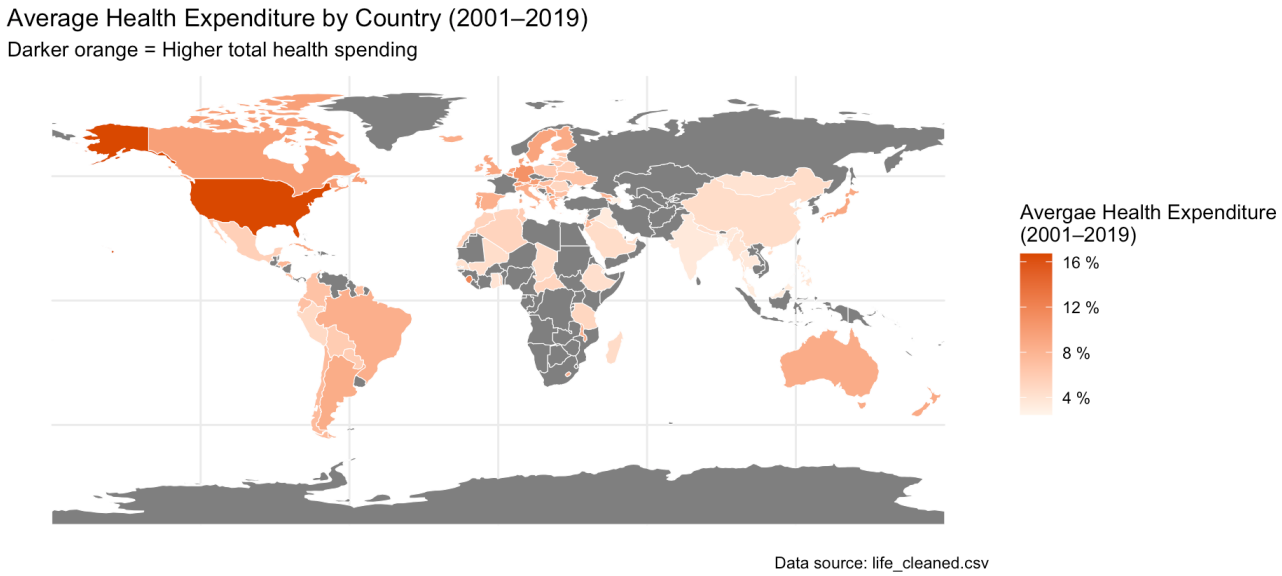
---

5. Final Model-MAP Model: Average Health Expenditure by Country (2001–2019)

To explore patterns in national health investments, we developed a map-based model showing total health expenditure from 2001 to 2019 across countries.

[Data-visualization-7: map model of Health.Expenditure vs Country Name, Author=Ruijie Li]

```
# Step 1: Calculate mean health expenditure per country from 2001 to 2019
health_total <- data %>%
  filter(Year >= 2001, Year <= 2019) %>%
  group_by(Country.Code, Country.Name) %>%
  summarise(Avg_Health_Spending = mean(Health.Expenditure.., na.rm = TRUE), .gro
ups = "drop")

# Step 2: Load world map
world <- ne_countries(returnclass = "sf")

# Step 3: Merge map data with health expenditure
map_data <- left_join(world, health_total, by = c("iso_a3" = "Country.Code"))

# Step 4: Plot the map (orange gradient)
ggplot(map_data) +
  geom_sf(aes(fill = Avg_Health_Spending), color = "white", size = 0.1) +
  scale_fill_gradient(
    low = "#fff5eb",
    high = "#d94801",
    name = "Avergae Health Expenditure\n(2001–2019)",
    labels = label_number(suffix = " %")
  ) +
  labs(
    title = "Average Health Expenditure by Country (2001–2019)",
    subtitle = "Darker orange = Higher total health spending",
    caption = "Data source: life_cleaned.csv"
  ) + theme_minimal()
```



Key Insights:

- Developed countries generally have higher total health expenditures, often driven by larger GDPs and more developed healthcare systems, for example Europe and Oceania.
- There is a clear regional disparity, highlighting gaps in healthcare funding and access between developed and developing countries.
- Most developing countries in south America has higher Health Expenditure and the USA has highest health expenditure but has lower health condition.

Conclusion: This global map underscores significant disparities in health investment. By comparing two map above, we can conclude that health indicators and health expenditures in most of countries are positively correlated. In Europe, Oceania and Latin America region, health expenditure is high and health indicator is also high. Middle East and North Africa areas has lower health expenditures but have higher health indicators. However, the USA has highest health expenditure but has lower health condition.

Final Answers to Research Questions

1. **Question 1**:
    - Health disparities between countries are primarily driven by differences in life expectancy, health expenditure, sanitation, and nutrition.
2. **Question 2**:
    - Average Health Expenditure (% of GDP) and Average Sanitation(access by % of population) both show a positive relationship with Life Expectancy (year), with correlation coefficients of r = 0.413 and r = 0.697, respectively. This reflects that it is very helpful to invest and improve health expenditure and sanitation to increase people's life expectancy.
3. **Question 3**:
    - The health indicators and health expenditures are positively correlated. In the most of countries, the higher health expenditure spend in total GDP, the lower health burden will have, means the health condition is better.

Discussion

1. **Impact of Socio-Economic and Environmental Factors on Life Expectancy**

- Government Health Expenditure: Our exploratory analysis indicates a significant positive correlation between government health expenditure and life expectancy across countries. The regression analysis shows that countries with higher health spending tend to have higher life expectancy (e.g., the correlation coefficient is high and the regression coefficient is significant with $p < 0.001$), suggesting that increased health investment contributes to improvements in public health.
- CO_2 Emissions: The scatter plot analysis reveals that higher CO_2 emissions are generally associated with lower life expectancy. This finding indicates that environmental pollution may negatively impact public health, highlighting the need for countries to address environmental

management alongside economic growth to improve health outcomes.

- Income Levels: The box plot analysis demonstrates clear differences in life expectancy across different income groups. High-income countries generally exhibit higher life expectancy compared to low-income countries, further confirming the crucial role of economic strength in public health.

2. **The regression analysis Between Life Expectancy, Sanitation, and Health Expenditure**

   The regression visualization above explore the relationships between life expectancy and access to sanitation and health expenditure across countries from 2001 to 2019.

- The first plot shows a moderate but positive association between **health expenditure (% of GDP)** and **life expectancy**, with a correlation coefficient of r = 0.413. While increased health spending has a positive influence on health outcomes, the relatively weak correlation reflects that Increasing health care costs may have a limited improvement on people's life expectancy. Notably, the United States is a special case. Although it spends significantly more on health care than other countries, its national life expectancy is comparable to that of countries that spend around 10% of their GDP on health.
- The second plot reveals a strong positive correlation between **sanitation access rate(%)** and **life expectancy** (r = 0.697). Countries with higher sanitation coverage tend to have longer life expectancy. All linear regression, LOESS, and GAM trends suggests a stable upward pattern across the full range of data. This shows that popularize and improve basic sanitation infrastructure can have significant benefits for population lifespans.

3. **PCA and Composite Health Indicators Analysis**

- PCA Analysis: We conducted a Principal Component Analysis (PCA) on average **life expectancy**, **health expenditure**, **sanitation conditions** (e.g., the percentage of the population using safely managed sanitation services), and **undernourishment rate** across countries.
- The results indicate that the first principal component (PC1) mainly reflects health investment and public health conditions, showing a positive relationship with life expectancy; the second principal component (PC2) may capture additional factors such as education expenditure or social welfare levels. By applying clustering methods (e.g., K-means clustering to divide countries into three groups), we can preliminarily categorize nations into groups with good, moderate, and poor health outcomes.

4. **Global Health Map Analysis**

- Map Model 1: The global health map visualizes the inverted **average burden of health loss** (a composite measure derived from **injuries**, **communicable**, and **non-communicable diseases**) for the period 2001 to 2019. On the map, darker green regions represent countries with lower inverted average burdens, which mean has better health condition, while lighter areas indicate higher health burdens.
- Map Model 2: The average health expenditure map represents the percentage of **health expenditure** in total GDP of each country. Darker orange regions represent countries with higher percentage in health expenditure, while lighter areas indicate lower spend in health.
- From these two map above, we can concluded that Europe, Latin America & Caribbean and Oceania regions have higher percentage in health expenditure and accordingly lower burden of

health loss. In the most of countries the health indicators and health expenditures are positively correlated. However, the USA has the highest percentage, which is over 16%, in health expenditure but has lower health condition.

Conclusion

This study finds that higher **health expenditure**, better **sanitation access**, and **higher income levels** are all positively associated with **life expectancy**. While increased spending generally improves health outcomes, the effect varies across countries—highlighted by the U.S., where high spending doesn't guarantee better results. PCA and clustering analysis further helped classify countries based on composite health indicators, and global maps revealed that regions with higher health investment tend to have lower health burdens. These results underscore the importance of both economic and environmental investments in enhancing public health.

Future Research Directions:

1. Explore additional variables (e.g. air quality, working hours) to consider more factors may impact life expectancy
2. Temporal Analysis: Extend the analysis beyond 2019 to explore evolving trends.
3. Advanced Modeling: Apply causal inference methods to better establish the impact of policy interventions on life expectancy.

References

1. Soumikniloy, "Life expectancy & Socio-Economic (world bank)," Kaggle.com, Sep. 13, 2023. https://www.kaggle.com/code/soumikniloy/life-expectancy-socio-economic-world-bank (https://www.kaggle.com/code/soumikniloy/life-expectancy-socio-economic-world-bank).

Team work

Zezhou Lu identified and selected the dataset used in this project.

Zezhou Lu and Ruijie Li shared and contributed their ideas to explore the research direction.

Exploratory Data Analysis (EDA) model:

boxplot(Life Expectancy vs. Income Group), scatter plot(Average Life Expectancy vs. Average CO2 Emission (2001-2019)) Histogram of Life Expectancyare made by Junyuan Wu.

Data Analysis:

Zezhou Lu was responsible for making and analyzing regression models of Average Life Expectancy vs. Average Health Expenditure (2001–2019) and Average Life Expectancy vs. Average Sanitation (2001–2019) between different countries. PCA + Kmeans (Average Health and Sanitation Conditions for Different Countries (2001–2019))

While Ruijie Li was responsible for making and analyzing maps which are the Global Health Map of Nagative Avg Inverted Burden (2001-2019) part and the Total Health Expenditure by Country.

In terms of report writing, Junyuan was responsible for Introduction and part of Data. Ruijie Li was responsible for Data, the part of research questions and final answers, reference and notes for plots.

Zezhou Lu was responsible for the conclusion, Teamwork, and part of the research questions and final answers. For the remaining part of report editing, such as discussion and so on, all members evenly distributed the workload.

Regarding the oral presentation and slide making, all members planned to present their own parts of data visualizations and finds.All members planned to distribute the workload evenly for common sharing parts, such as introduction and conclusion.

Zezhou Lu and Ruijie Li reviewed the progress, edited and revised the visualizations and text.

Notes

Datasets: there exists **world bank group** to the `Life expectancy & Socio-Economic` dataset, featuring recent data with similar metrics (country, years, Life expectancy)

For examples of data visualizations, see: https://data.worldbank.org/ (https://data.worldbank.org/)