

Mini Project 1

PSTAT100: Data Science Concepts and Analysis

Instructor: Ali Abuzaid


2024-10-22

STUDENT NAME

- Yue Zhu A1F8C52
- Harry He 7133085
- Tina Zhou 4876165
- Xihua Yao 5927777

Instructions

- This mini project aims to familiarize you with real-life data sourced from various resources.
- The mini project includes narrative questions. While these questions are primarily based on lecture material and prerequisites, they may also require independent thinking and investigation.
- Collaborate in groups of **3-4** students from the **same discussion session**; individual submissions **will not be accepted**.
- Submit the the answers at **Gradescope**.
- Please use the provided `.qmd` file to type your solutions and submit the completed project as a PDF file. You can utilize **RStudio** for this purpose. For guidance, refer to the [Tutorial: Hello, Quarto](#)).
- Ensure that all **R** code, mathematical formulas, and workings are presented clearly and appropriately.
- All figures should be numbered, and axes must be labeled.

 Due Date

Due Date: Friday, October 18, 2024, 11:59 PM

1 Introduction

This comprehensive dataset serves as a vital resource for understanding global health dynamics and demographics, providing insights into life expectancy and population health across diverse nations. By examining key indicators such as adult mortality, infant deaths, and health expenditures, researchers can uncover critical trends and factors influencing public health outcomes. Sourced from reputable global health organizations, this data is essential for informing health policy and advancing epidemiological research.

1.1 Data Source:

Laksika Tharmalingam (2023), “Health and Demographics Dataset” Retrieved [September 29, 2024] from (<https://www.kaggle.com/datasets/uom190346a/health-and-demographics-dataset>)

1.2 Data Description

We will refer to the data as **Health**. The **Health** dataset contains of the following variables.

Variable	Description
Country	Explore the global tapestry with data from diverse nations.
Year	Unlock the passage of time and its impact on health trends.
Status	Understand the development status, whether “Developed” or “Developing,” that shapes the course of health.
Life Expectancy	Peer into the crystal ball of population health, revealing how long people can expect to live.
Adult Mortality	Gauge the probabilities of survival between ages 15 and 60 per 1,000 population.
Infant Deaths	Delve into infant health with the number of infant deaths per 1,000 live births.
Alcohol	Raise a glass to insights on average alcohol consumption in liters per capita.
Percentage Expenditure	Unearth health expenditure as a percentage of a country’s GDP.
Hepatitis B	Measure immunization coverage for Hepatitis B.
Measles	Examine the impact of this preventable disease with the number of reported cases per 1,000 population.
BMI	Step onto the scales of national health with the average Body Mass Index.
Under-Five Deaths	Shine a spotlight on child mortality with the number of deaths under age five per 1,000 live births.

Variable	Description
Polio	Inspect immunization coverage for Polio.
Total Expenditure	Track the total health expenditure as a percentage of GDP.
Diphtheria	Assess immunization coverage for Diphtheria.
HIV/AIDS	Witness the prevalence of HIV/AIDS as a percentage of the population.
GDP	Follow the financial pulse of a nation with Gross Domestic Product data.
Population	Witness the ebb and flow of a nation's populace.
Thinness 1-19 Years	Explore the prevalence of thinness among children and adolescents aged 1-19.
Thinness 5-9 Years	Zoom in on thinness among children aged 5-9.
Income Composition of Resources	Decode the composite index reflecting income distribution and resource access.
Schooling	Measure the gift of knowledge with data on average years of schooling.

2 Project Questions

ANSWER ALL THE FOLLOWING QUESTIONS:

! Question 1

1. How can you classify the **Health** dataset: is it structured or unstructured?
2. What are the observational units in the **Health** dataset?
3. What is the number of observations?
4. What is the number of variables?
5. For each of the following variables, specify its type (quantitative or qualitative). If it is quantitative, indicate whether it is discrete or continuous, and describe its scale of measurement.

1. The dataset is structured.
2. The observational units in the Health dataset are countries observed over different years.
3. 1649
4. 22

5.

Variable	Variable Type	Quantitative Type	Scale
Country	Qualitative	/	Nominal scale
Year	Quantitative	Discrete	Interval scale
Status	Qualitative	/	Nominal scale
Life Expectancy	Quantitative	Continuous	Ratio scale
Adult Mortality	Quantitative	Discrete	Ratio scale
Infant Deaths	Quantitative	Discrete	Ratio scale
Alcohol	Quantitative	Continuous	Ratio scale
Percentage Expenditure	Quantitative	Continuous	Ratio scale
Hepatitis B	Quantitative	Discrete	Ratio scale
Measles	Quantitative	Discrete	Ratio scale
BMI	Quantitative	Continuous	Ratio scale
Under-Five Deaths	Quantitative	Discrete	Ratio scale
Polio	Quantitative	Discrete	Ratio scale
Total Expenditure	Quantitative	Continuous	Ratio scale
Diphtheria	Quantitative	Discrete	Ratio scale
HIV/AIDS	Quantitative	Continuous	Ratio scale
GDP	Quantitative	Continuous	Ratio scale
Population	Quantitative	Continuous	Ratio scale
Thinness 1-19 Years	Quantitative	Continuous	Ratio scale
Thinness 5-9 Years	Quantitative	Continuous	Ratio scale
Income Composition of Resources	Quantitative	Continuous	Ratio scale
Schooling	Quantitative	Continuous	Ratio scale

! Question 2

Every country has different observations that vary across different years, while its classification as “Developed” or “Developing” generally remains consistent.

- a. Arrange the `Health` dataset so that each country is classified according to its development status and the number of years represented.
- b. Comment on the inclusion of countries, their development status, and the period of registration.

ANSWERS TO QUESTION 2:

```
1 # (a)
2 library(readxl)
3 health_data <- read_excel("Health Data.xlsx")
4 library(tidyr)
5 library(dplyr)
6 health_data_long <- health_data[c("Country", "Status", "Year")] %>% group_by(Country, Status)
7 print(health_data_long)
```

```
# A tibble: 133 x 3
# Groups:   Country, Status [133]
  Country      Status      n
  <chr>        <chr>    <int>
1 Afghanistan Developing    16
2 Albania      Developing    16
3 Algeria      Developing    11
4 Angola       Developing     8
5 Argentina    Developing    13
6 Armenia      Developing    15
7 Australia    Developed     14
8 Austria      Developed     15
9 Azerbaijan   Developing    13
10 Bangladesh  Developing    12
# i 123 more rows
```

#(b) There are notable differences in the inclusion of countries, their development status, and the period of data registration. These countries are classified as either “Developed” or “Developing,” which generally reflects the global distribution of economic and social development levels. While many countries have relatively complete data, some countries have shorter

periods of data, possibly due to limitations in data collection. Overall, the dataset achieves a reasonable balance between developed and developing countries, but there may be some bias in terms of time coverage and country inclusion, which requires further analysis to ensure representativeness and accuracy in classification.

! Question 3

- calculate the means of all other numeric variables for each country while excluding the “Status” variable.
- Visualize the distribution of the obtained means for Polio, Life expectancy and BMI. Comment on the distribution of the three variables.

ANSWERS TO QUESTION 3:

Replace this line with your answers

```
1 #a)
2 library(dplyr)
3
4 numeric_vars <- health_data %>%
5   select(-Status) %>%           # Exclude the 'Status' column
6   group_by(Country) %>%         # Group by 'Country'
7   summarise(across(where(is.numeric), mean, na.rm = TRUE))
8
9 head(numeric_vars)
```



```
# A tibble: 6 x 21
  Country      Year `Life expectancy` `Adult Mortality` `infant deaths` Alcohol
  <chr>      <dbl>         <dbl>           <dbl>           <dbl>    <dbl>
1 Afghanistan 2008.         58.2             269.             78.2    0.0144
2 Albania     2008.         75.2             45.1             0.688    4.85
3 Algeria     2009          74.2            103.             20.3    0.447
4 Angola      2010.         50.7            363.             76.6    7.62
5 Argentina   2008          75.2            100.             10      8.00
6 Armenia     2007          73.3            117.              1      3.70
# i 15 more variables: `percentage expenditure` <dbl>, `Hepatitis B` <dbl>,
#   Measles <dbl>, BMI <dbl>, `under-five deaths` <dbl>, Polio <dbl>,
#   `Total expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>, GDP <dbl>,
#   Population <dbl>, `thinness 1-19 years` <dbl>, `thinness 5-9 years` <dbl>,
#   `Income composition of resources` <dbl>, Schooling <dbl>
```



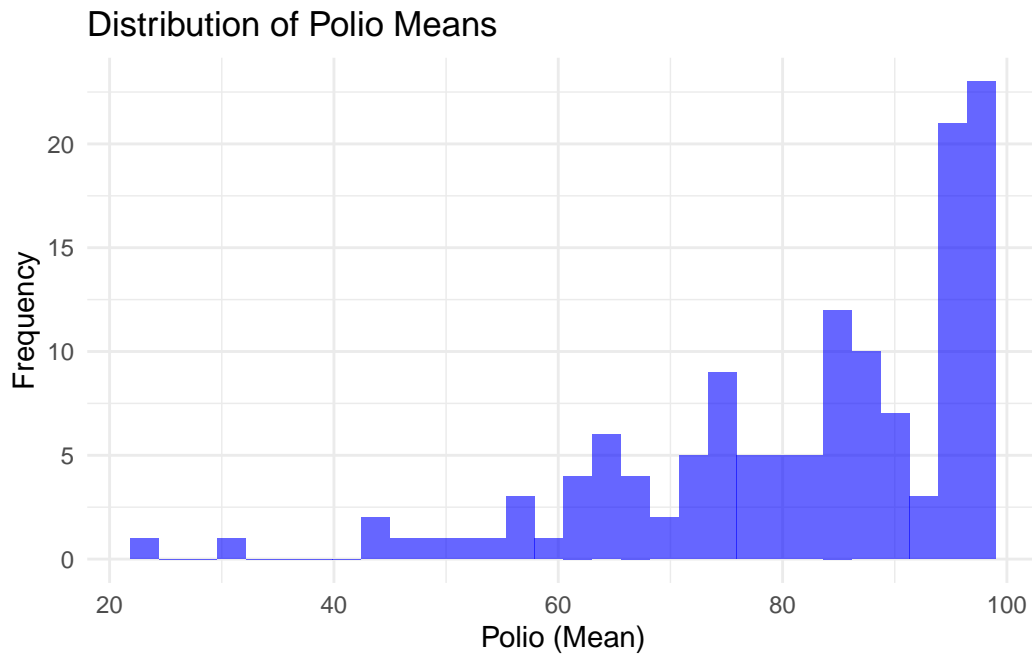
```
1 #b)
2
3 library(ggplot2)
```



```

4
5 ggplot(numeric_vars, aes(x = Polio)) +
6   geom_histogram(bins = 30, fill = "blue", alpha = 0.6) +
7   labs(title = "Distribution of Polio Means", x = "Polio (Mean)", y = "Frequency") +
8   theme_minimal()

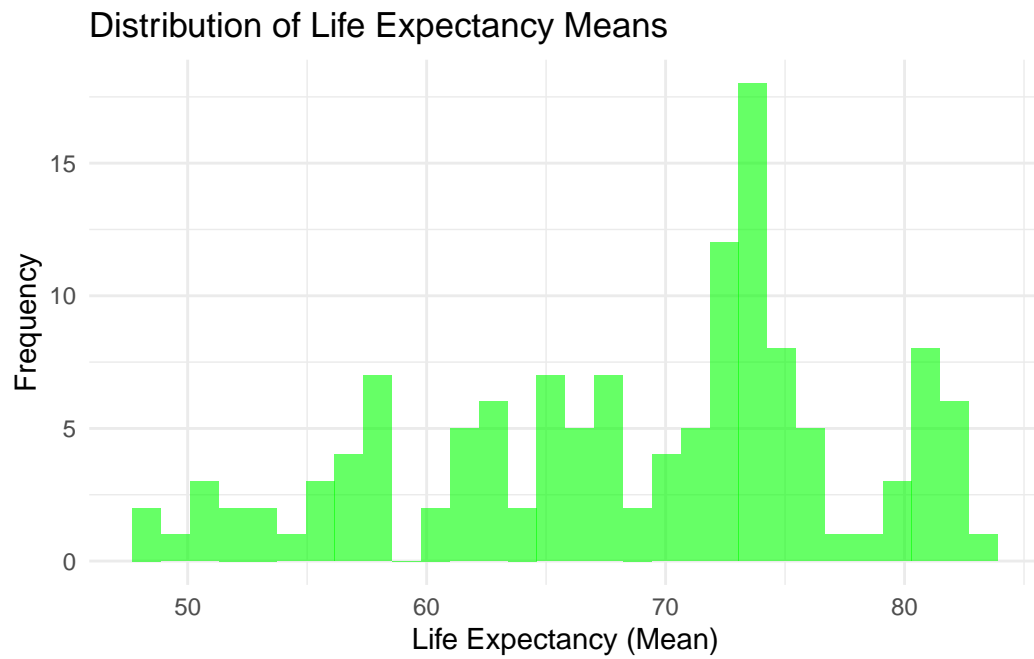
```



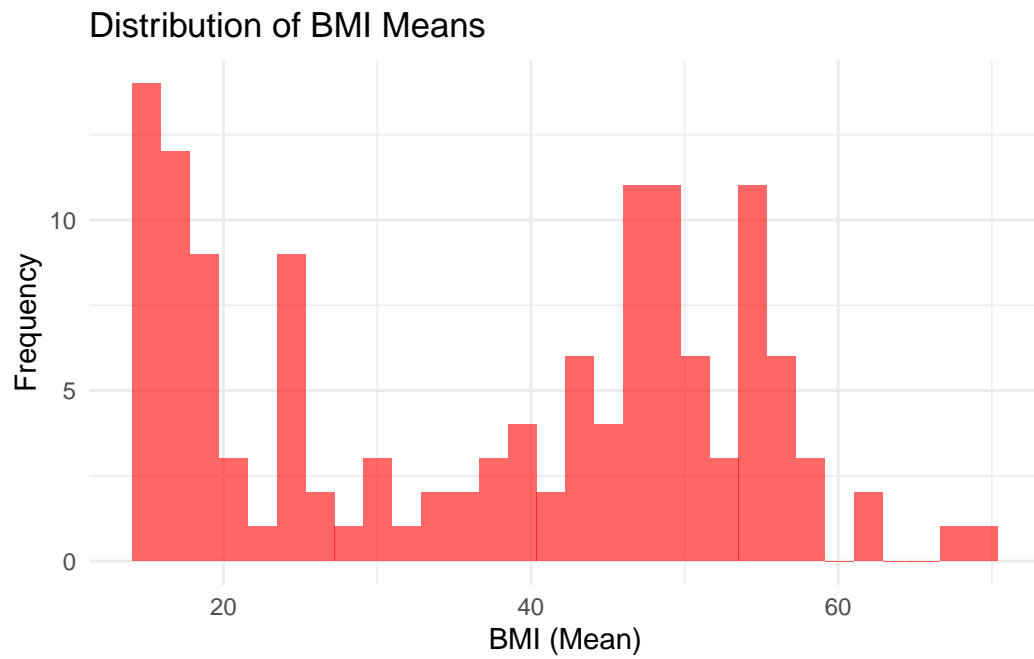
```

1 ggplot(numeric_vars, aes(x = `Life expectancy`)) +
2   geom_histogram(bins = 30, fill = "green", alpha = 0.6) +
3   labs(title = "Distribution of Life Expectancy Means", x = "Life Expectancy (Mean)", y =
4   theme_minimal()

```



```
1 ggplot(numeric_vars, aes(x = BMI)) +  
2   geom_histogram(bins = 30, fill = "red", alpha = 0.6) +  
3   labs(title = "Distribution of BMI Means", x = "BMI (Mean)", y = "Frequency") +  
4   theme_minimal()
```



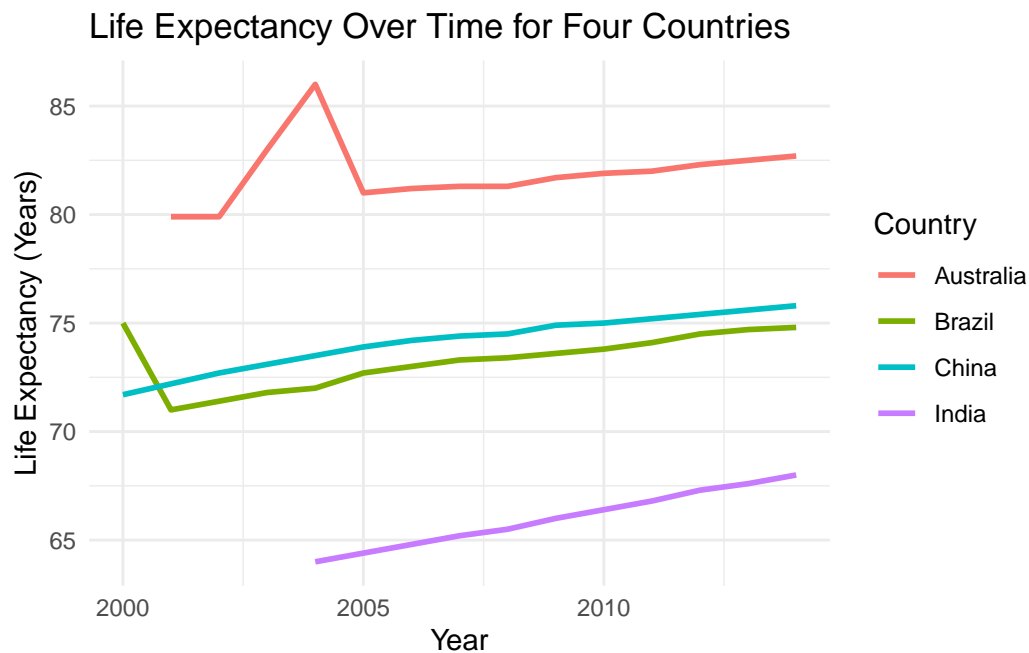
```
1 # For polio, the distribution is left-skewed, with a peak between 95 and 100. There is also a smaller peak around 105.
2 # The life expectancy distribution is slightly left-skewed and close to a normal distribution.
3 # The histogram of BMI means is bimodal, with two clear peaks. The distribution is concentrated between 10 and 70.
```

! Question 4

Create a time series plot for life expectancy in four countries over the given years.
Comment on the plot.

ANSWERS TO QUESTION 4:

```
1 library(ggplot2)
2 library(dplyr)
3 library(readxl)
4 selected_countries <- filter(health_data, Country %in% c("Australia", "China", "India", "B
5 ggplot(selected_countries, aes(x = Year, y = `Life expectancy`, color = Country)) +
6   geom_line(size = 1) +
7   labs(title = "Life Expectancy Over Time for Four Countries",
8         x = "Year",
9         y = "Life Expectancy (Years)") +
10  theme_minimal()
```



Australia (Red Line):

Australia's life expectancy shows a sharp peak around 2004, reaching about 85 years, followed

by a slight decrease while still maintaining a high level.

Brazil (Green Line):

Brazil's life expectancy starts at around 75 years, decreases to 71 then shows a relatively steady upward trend toward 75.

China (Blue Line):

China's life expectancy starts at around 72 years and steadily increases to about 76 years.

India (Purple Line):

Starting from 2004, India has the lowest starting life expectancy, around 64 years, but shows a notable upward trend.

Trends:

Australia, as a developed country, has a much higher life expectancy than Brazil, China, and India, though it shows some fluctuations. China, India and Brazil are growing fast, reflecting improvements in their economies and healthcare, while China and India has a more stable and gradual increase.

Development Differences:

The graph shows a clear gap in life expectancy between developed countries like Australia and developing countries like Brazil, China, and India.

! Question 5

- a. How might geographical location affect life expectancy? Can we visualize the mean life expectancy in an appropriate way to evaluate this hypothesis?
- b. What insights can be drawn from the constructed graph, and how does incomplete data impact our understanding of the results?

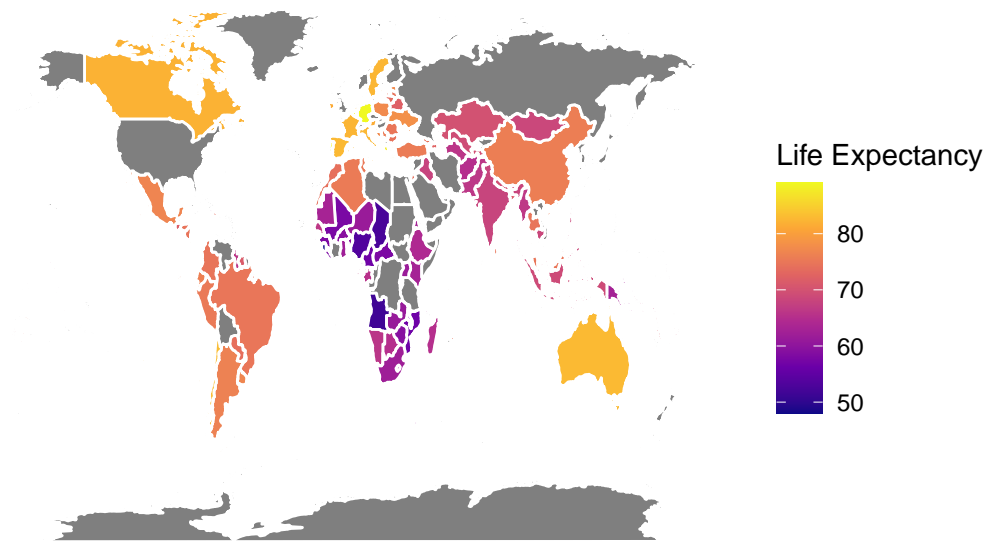
3 Hint

One can use the `ggplot2` and `maps` libraries in R. Use the function `left_join` to match `map_data` to `Country` in `Health` dataset.

ANSWERS TO QUESTION 5:

```
1 world_map <- map_data("world")
2 health_data_clean <- health_data %>%
3   distinct(Country, .keep_all = TRUE)
4 health_map_data <- left_join(world_map, health_data_clean, by = c("region" = "Country"))
5 ggplot(health_map_data, aes(long, lat, group = group, fill = `Life expectancy`)) +
6   geom_polygon(color = "white") +
7   scale_fill_viridis_c(option = "C", na.value = "gray50") +
8   labs(title = "Geographical Influence on Life Expectancy",
9        fill = "Life Expectancy") +
10  theme_minimal() +
11  theme(axis.text = element_blank(),
12        axis.title = element_blank(),
13        panel.grid = element_blank())
```

Geographical Influence on Life Expectancy



Life expectancy is influenced by several key factors. GDP reflects a country's economic status and access to healthcare and resources, playing a significant role in determining life expectancy. Education also has a strong impact, as higher levels of education are associated with greater health awareness and healthier lifestyle choices. Furthermore, healthcare expenditure indicates the level of investment in public health. Lifestyle factors such as alcohol consumption and BMI are also important, with excessive alcohol use and higher BMI often linked to shorter lifespans.

Infant mortality rates show significant regional disparities, with developed countries typically having much lower rates than their developing counterparts. In regions like Sub-Saharan Africa and South Asia, infant mortality remains high, driven by factors such as limited healthcare access, inadequate nutrition, and a greater prevalence of infectious diseases. Conversely, regions such as North America and Europe benefit from lower infant mortality rates, thanks to well-established healthcare systems and robust public health programs.

! Question 6

Discuss the following questions:

- What factors (e.g., alcohol consumption, GDP, education) most strongly correlate with life expectancy?.
- Are there any notable differences in infant death rates between regions?

ANSWERS TO QUESTION 6:

a.

```
1 life_expectancy <- health_data$`Life expectancy`  
2 #cor(alcohol_consumption, life_expectancy, method = "pearson")  
3 cor(health_data[,c(5:22)], life_expectancy, method = "pearson")
```

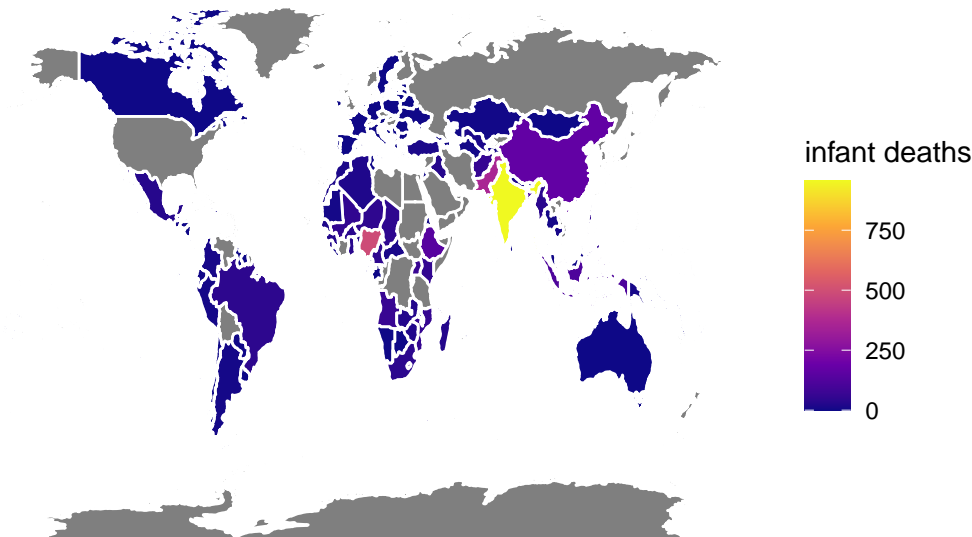
	[,1]
Adult Mortality	-0.70252306
infant deaths	-0.16907380
Alcohol	0.40271832
percentage expenditure	0.40963082
Hepatitis B	0.19993528
Measles	-0.06888122
BMI	0.54204159
under-five deaths	-0.19226530
Polio	0.32729440
Total expenditure	0.17471764
Diphtheria	0.34133123
HIV/AIDS	-0.59223629
GDP	0.44132181
Population	-0.02230498
thinness 1-19 years	-0.45783819
thinness 5-9 years	-0.45750829
Income composition of resources	0.72108259
Schooling	0.72763003

- The correlation analysis reveals that schooling (0.73) and income composition (0.72) have the strongest positive relationships with life expectancy, indicating that better educational opportunities and equitable wealth distribution lead to healthier populations. GDP also shows a moderate positive correlation (0.44), suggesting that economic prosperity contributes to better living conditions and access to healthcare, which in turn

enhances life expectancy. Additionally, moderate alcohol consumption shows a positive relationship (0.40), potentially reflecting the health benefits of moderate drinking, though excessive alcohol use can have detrimental effects. In contrast, higher rates of adult mortality (-0.70) and the prevalence of diseases such as HIV/AIDS (-0.59) negatively impact life expectancy, highlighting the role of health interventions and disease management in improving lifespan. Thus, a combination of economic stability, education, and effective health policies is essential for enhancing life expectancy across populations.

```
1 world_map <- map_data("world")
2 health_data_clean <- health_data %>%
3   distinct(Country, .keep_all = TRUE)
4 health_map_data <- left_join(world_map, health_data_clean, by = c("region" = "Country"))
5 ggplot(health_map_data, aes(long, lat, group = group, fill = `infant deaths`)) +
6   geom_polygon(color = "white") +
7   scale_fill_viridis_c(option = "C", na.value = "gray50") +
8   labs(title = "Different Infant Death Rates Between Regions",
9        fill = "infant deaths") +
10  theme_minimal() +
11  theme(axis.text = element_blank(),
12        axis.title = element_blank(),
13        panel.grid = element_blank())
```

Different Infant Death Rates Between Regions



b. Infant mortality rates often vary significantly between different regions due to disparities

in healthcare access, economic conditions, and public health interventions. Developed countries like Australia, Canada and European countries typically have lower infant death rates due to advanced healthcare systems, widespread immunization, and better maternal care. Developing countries, especially India, which has the highest infant death rate, may struggle with higher infant mortality rates due to challenges like malnutrition, poor access to healthcare, and preventable diseases.