# Country Health Index (CHI)

# Data Analysis and Visualization Documentation

CA1 Index Generation and Visualization **Matthew Riddell - D00245674**

**GitHub Repo:** https://github.com/Matthew-Riddell/DAV_CA1

**ChatGPT Chat Log:** https://chatgpt.com/share/68166c24-14cc-8007-a343-b1b7946badd2

**Dataset:** https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023

## 1. Theoretical Framework

The Country Healthy Index (CHI) is a composite Index designed to measure the health of a given country based on an assessment of various health, social and environment indicators to get a comprehensive view of the country's medical and health landscape, after assessment of each variable the CDI then provides a ranking between 0-100 with 0 being the least healthiest country and 100 being the healthiest country. By focusing on these various aspects of the country's society, healthcare system, environmental landscape, this index aims to highlight not only direct health outcomes among the population but also underlaying systemic factors that may contribute the country's overall health.

**Dimensions:**

The Country Healthy Index (CHI) employs the following Dimensions to seek a comprehensive view of the country's health:

**Longevity:**

Longevity refers to the average overall life expectancy of a country's population, it provides a simple metric for how long a person is expected to live in each society. Countries with higher life expectancies generally have better access to healthcare, better healthcare quality, better socioeconomic conditions, healthier lifestyles etc. The longevity dimension will be sourced from life expectancy data.

- **Theoretical Basis**: Longevity is influenced by various factors including access to healthcare, lifestyle habits, environmental influences, social economic conditions etc. all of which influence a country's overall health

**Access to Healthcare:**

Access to healthcare measures how easy it is for a countries population to receive healthcare. This includes availability of healthcare facilities, access to doctors, density of patients to doctors etc. Countries with a higher access to health care will generally have better health outcomes as patients receive the care, they need without any major barriers, saving lives.

- **Theoretical Basis:** According to the World Health Organization (WHO), access to healthcare is a critical determiner of health equity. It affects how well people can manager their health, avoid preventable diseases, receive treatment on time. While poor access to healthcare shows disparities in healthcare outcomes and greater mortalities **(World Health Organization, 2025)**.

**Health Outcomes:**

Health outcomes are measures that directly reflect the physical wellbeing of a country's population. These can include metrics such as infant mortality rates, maternal mortality rates, morbidity rates etc. Positive Health Outcomes are generally associated with higher standards of healthcare, healthier lifestyles and better environmental conditions in the given country.

- **Theoretical Basis:** Health outcomes are a direct result of the interactions between healthcare systems, behaviour, environmental factors, genetics etc. Health disparities often arise in these areas and outcomes serve as a reflection of the country's healthcare system efficiency and equity **(World Health Organization, 2025)**.

**Financial Risk:**

Financial risk refers to the economic burden of the healthcare in a country. This can include the cost of medical treatments, out of pocket expenditure for patients, etc. Countries with a high financial risk may face issues such as healthcare inequity where only the wealthiest can afford healthcare or economic instability which may result in reduced investment into healthcare.

- **Theoretical Basis:** The availability of funds for healthcare directly affects the quality and accessibility of healthcare services. Countries with higher healthcare costs or significant financial barriers to healthcare may have worse health outcomes due to uneven access to healthcare as only the richest can afford it leaving many without adequate healthcare.

**Environmental Factors:**

Environmental Factors encompass the physical and social environments in which a country's population lives. Air quality, water quality, sanitation, climate conditions etc. are all environmental factors that can play a role in the health of the people who live there. Air pollution can have negative health impacts and result in the prevalence of respiratory illnesses.

- **Theoretical Basis:** Environmental Factors play a critical role in public health, as living conditions, air quality, pollution, and sanitation directly affect the population as poor environmental and living conditions results in a reduced quality of life and a decline in health and an increase prevalence of diseases and illnesses. A healthier environment supports better quality of life and reduced healthcare needs.

**Data Source:**

The data for constructing the Country Health Index (CHI) will be sourced from the following dataset:

**Global Country Information Dataset 2023:**

The "Countries of the World 2023" provides global statistics on various health, economic, demographic factors, and other factors influencing health outcomes that will be instrumental in calculating each dimension of the CHI

**Source:**
[https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023](https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023)

# 2. Data Selection

For the development of the Country Healthy Index (CHI), I have chosen a set of variables from the "Countries of the World 2023" dataset that I believe provide a comprehensive view of the country's health status, while adhering to the above defined dimensions in the theoretical framework and they reflect various aspects such as healthcare outcomes, healthcare access, environmental factors, etc. Below are the specific variables I am using:

**Life Expectancy:**

Life expectancy refers to the average number of years that a newborn is expected to live. Life expectancy is one of the most common indicators of a country's overall health. It gives me a clear picture of how long people can expect to live, which is influenced by factors such as healthcare quality, lifestyle and socioeconomic conditions. Countries with higher life expectancies tend to have better healthcare systems and healthier populations so it's a key measure for my Index.

**Physicians per thousand:**

This variable measures the number of physicians (doctors) available per 1000 people in a country. Access to healthcare is a crucial factor with regards to health outcomes and numbers of physicians per 1000 people is a direct measure of healthcare accessibility and shows the density of doctor to patients. More physicians mean better access to care and less overall strain on the country's healthcare system, fewer physicians mean greater overall strain on the system and poorer access to healthcare, greater wait times etc. leading to a greater influence on health outcomes. This variable will help me determine how well distributed healthcare services are in each country.

**Infant Mortality:**

Infant morality refers to the number of deaths of infants under the age of 1 per 1000 live births. Infant mortality is a strong indicator of a country's healthcare system effectiveness, especially in relation to pregnancy health care and post pregnancy healthcare. A high infant mortality rate generally points to deficiencies in the country's healthcare system, poor environmental conditions, poor sanitation, poor nutrition etc. As such I believe this variable will provide significant insights into the quality of healthcare in each country.

**Maternal Mortality Ratio:**

This variable is a ratio between the number of maternal deaths per 100,000 live births due to complications caused by pregnancy or childbirth. Maternal Mortality is another important health indicator as it reflects the quality and accessibility to maternal healthcare. A high maternal mortality rate suggests inadequate access to maternal healthcare or poor-quality healthcare and services for maternity and pregnancies. This measure will allow me to assess how well a country supports the health of its mothers and the healthcare system's ability and capacity to support maternal health.

**Out of Pocket Health Expenditure:**

This variable measures the total percentage of total health expenditure paid out of the patient's own pocket rather than by the state, or through health insurance or other payouts/subsidies. The out-of-pocket health expenditure is an important measure of the financial burden faced by patients seeking healthcare. Countries with higher out of pocket expenses will generally have greater financial barriers to healthcare leading to delayed treatments or healthcare needs left unmet. This will help me asses how financially accessible the healthcare is for people in different countries.

**CO2 Emissions:**

This variable measures the amount of Carbon Dioxide (CO2) emissions produced by a country measured in tons. Environmental factors such as air quality and CO2 Emissions play a significant role in public health as higher CO2 emissions is generally a sign of greater pollution and linked to environmental damage. Poor air quality can cause many different respiratory diseases. By including CO2 emissions in my index, I can account for environmental health risks that may contribute to the health of a country.

# 3. Imputation of Missing Data

Upon importation of the dataset into the Jupyter notebook using the above variables that I defined to align with the dimensions of the theoretical framework, I found a lot of missing values and string variables that I needed to clean up and fix so that they could be used for the creating of the Country Health Index, below is the untouched display of some of the rows from the dataset:

| | Country | life_expectancy | physicians_per_thousand | infant_mortality | maternal_mortality_ratio | out_of_pocket_expenditure | co2_emissions |
|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 64.5 | 0.28 | 47.9 | 638.0 | 78.40% | 8,672 |
| 1 | Albania | 78.5 | 1.20 | 7.8 | 15.0 | 56.90% | 4,536 |
| 2 | Algeria | 76.7 | 1.72 | 20.1 | 112.0 | 28.10% | 150,006 |
| 3 | Andorra | NaN | 3.33 | 2.7 | NaN | 36.40% | 469 |
| 4 | Angola | 60.8 | 0.21 | 51.6 | 241.0 | 33.40% | 34,693 |

To do this, I added a print statement to count the number of instances where a variable has an empty entry using. isnull().sum.

```
Country                       0
life_expectancy               8
physicians_per_thousand       7
infant_mortality              6
maternal_mortality_ratio     14
out_of_pocket_expenditure     7
co2_emissions                 7
```

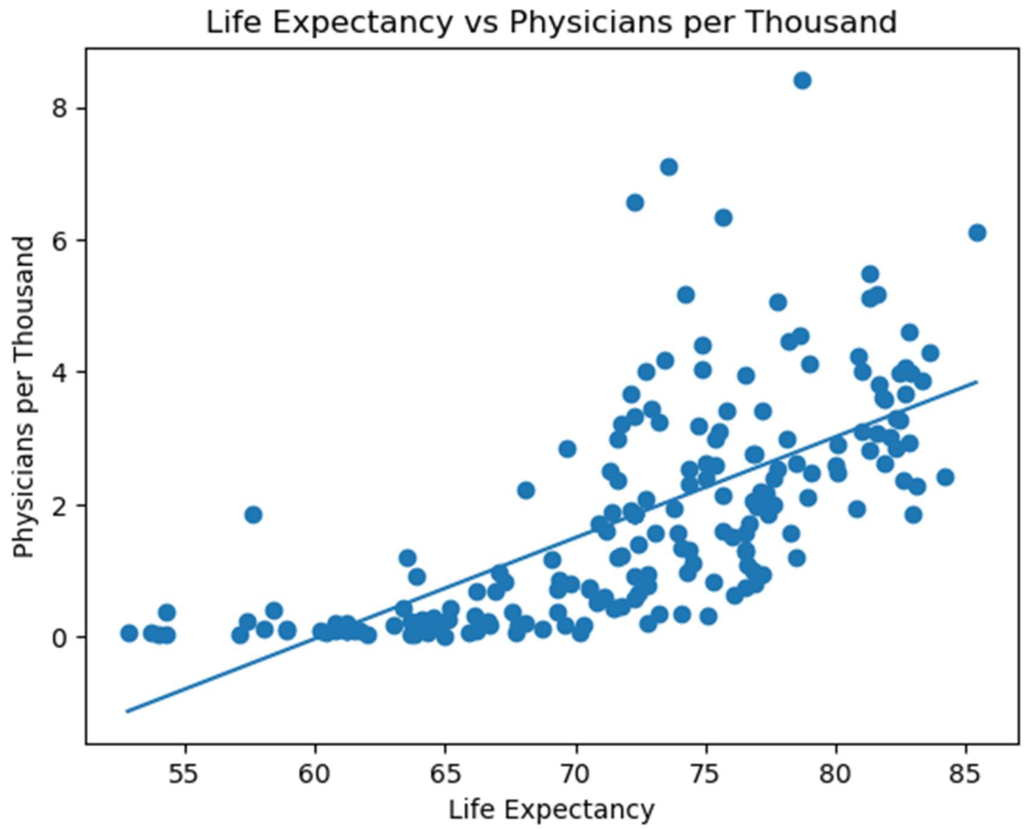
I then set out to strip the % from the out-of-pocket expenditure variable and convert them from string to float. I did the same thing with CO2 Emissions variable where I eliminated the comma and converted it to float. To address the missing values, I decided to insert mean averages of each column. I then did a second check to see if there were any remaining empty values:

```
Country                         0
life_expectancy                 0
physicians_per_thousand         0
infant_mortality                0
out_of_pocket_expenditure       0
co2_emissions                   0
country_healthy_index           0
country_healthy_index_scaled    0
```

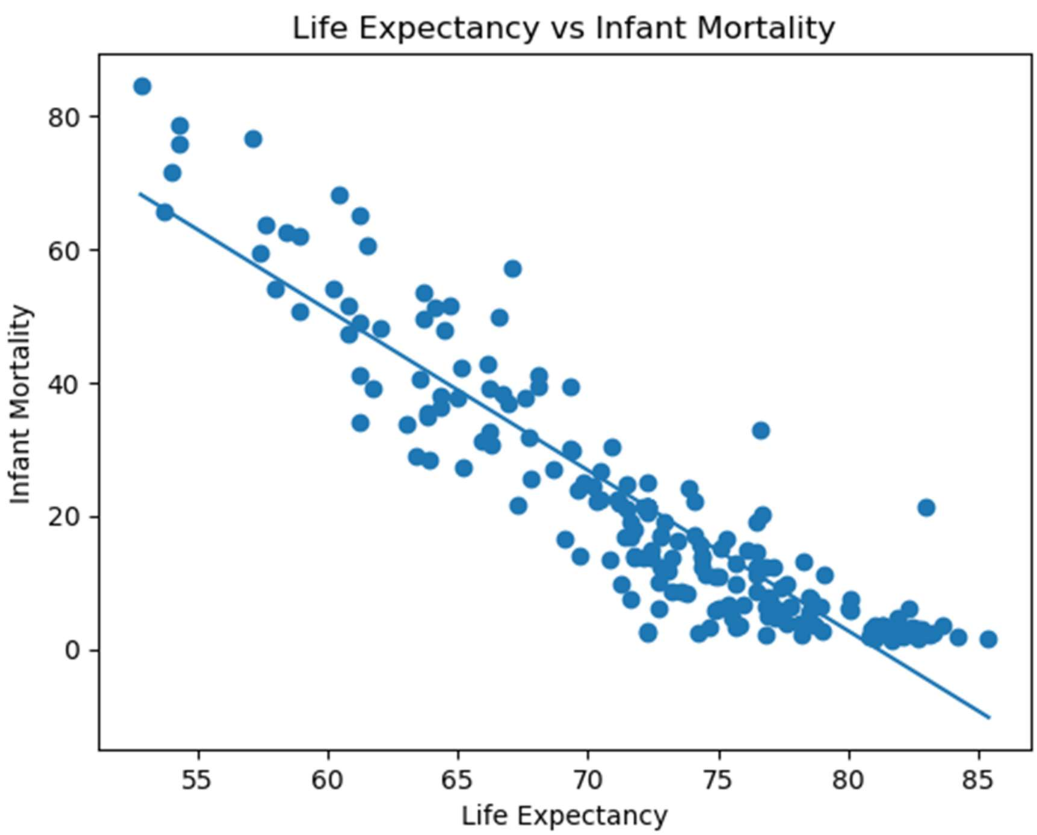
Here you can see that I no longer have missing values in my data and now I can proceed to Multivariate Analysis.

## 4. Multivariate Analysis

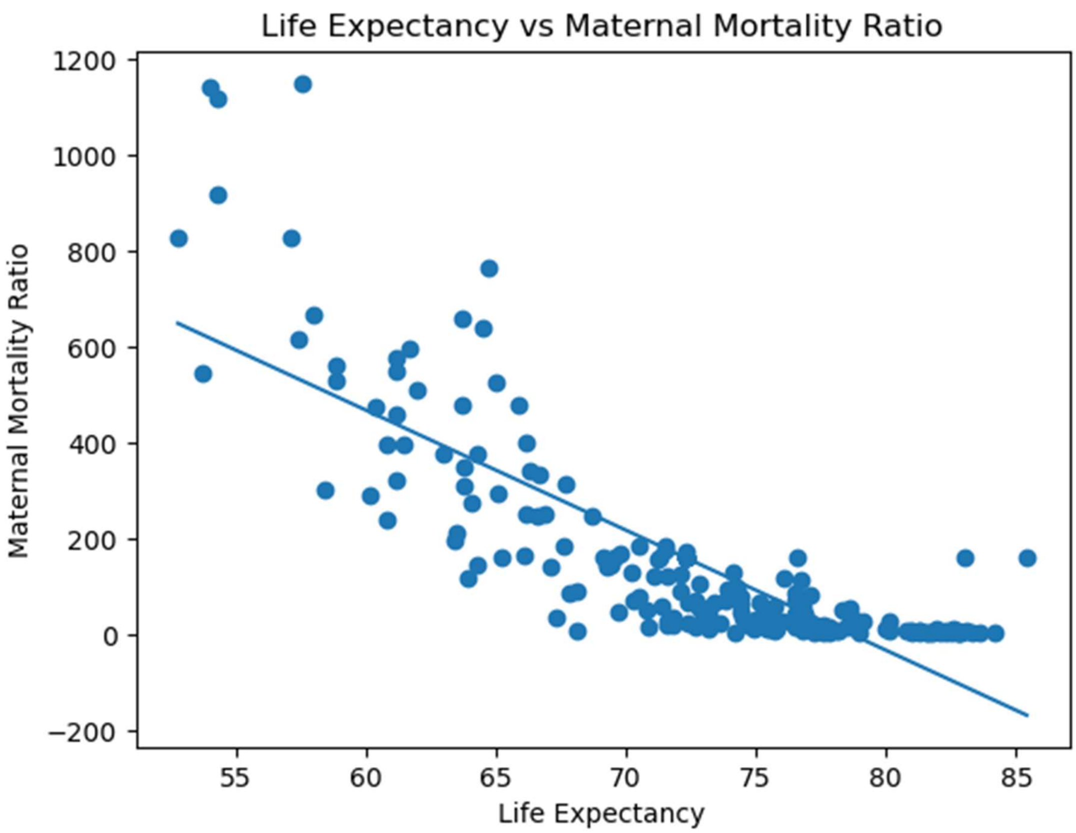
Once I have completely cleaned dataset, my next step is to perform multivariate analysis to better understand the relationships between each of the different variables that I have selected for the Country Health Index (CHI). To better visualise the relationships between the variables for Multivariate Analysis I decided to do correlation plots and matrices between all variables:

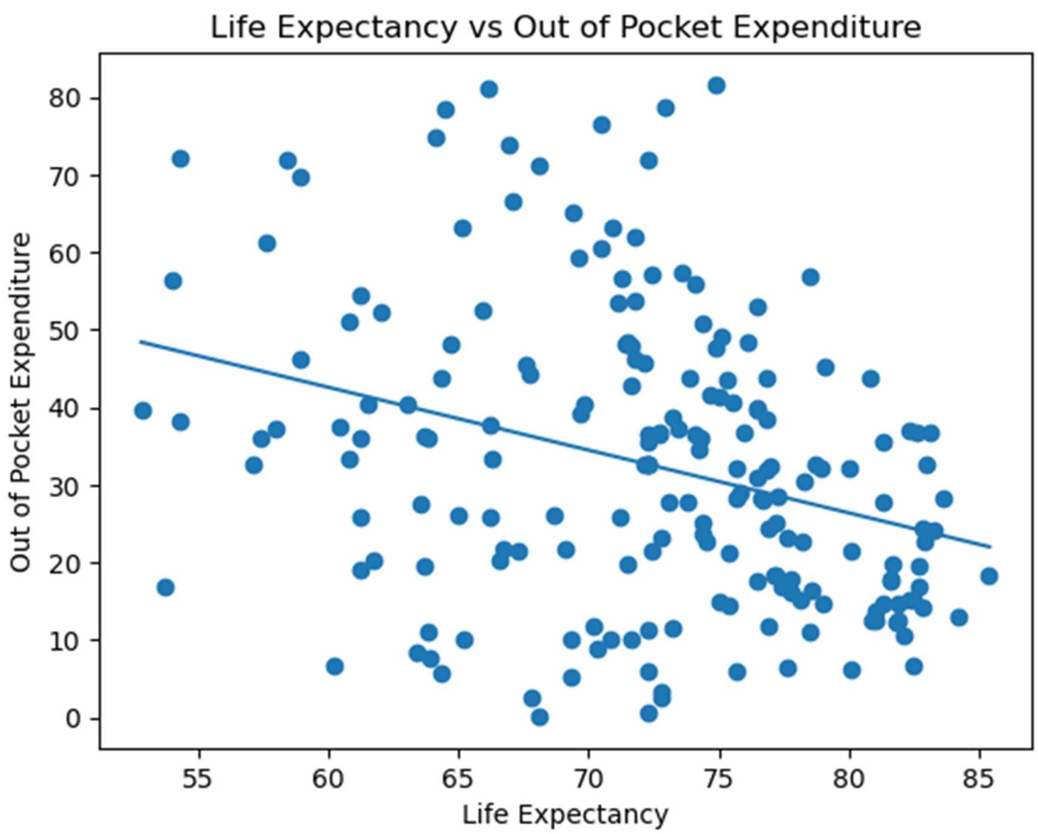


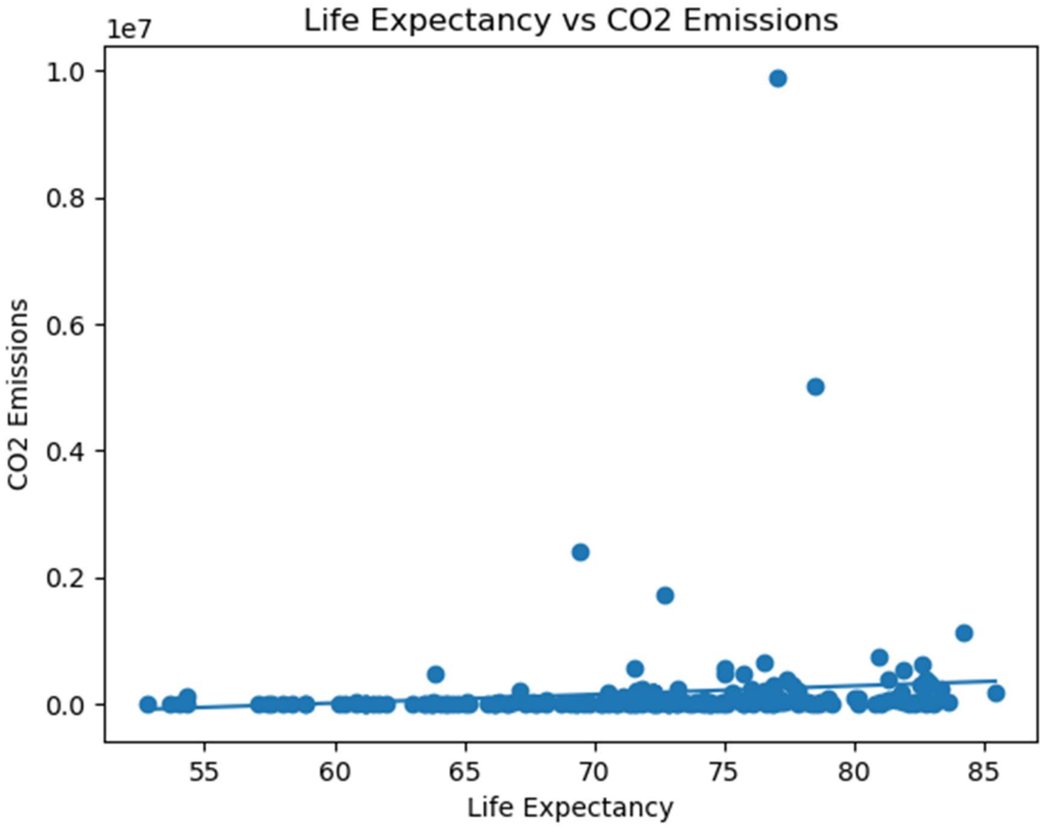
Life Expectancy vs Physicians per Thousand
Physicians per Thousand
Life Expectancy
0
2
4
6
8
55
60
65
70
75
80
85

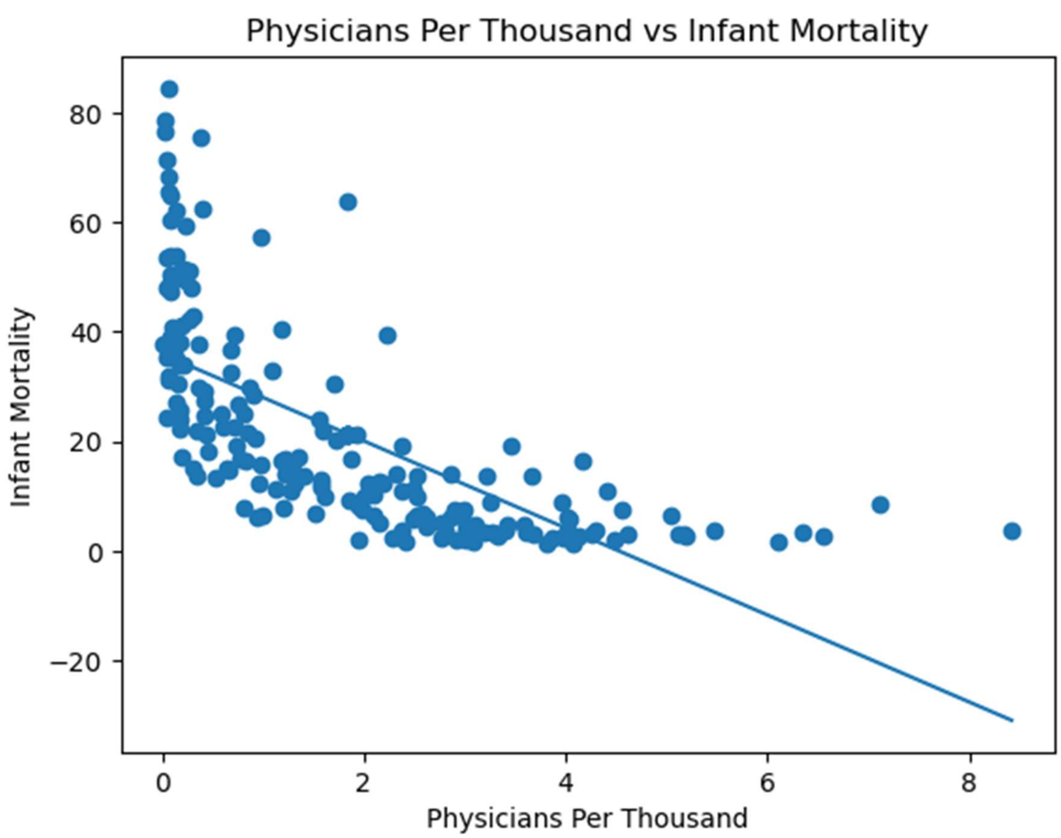



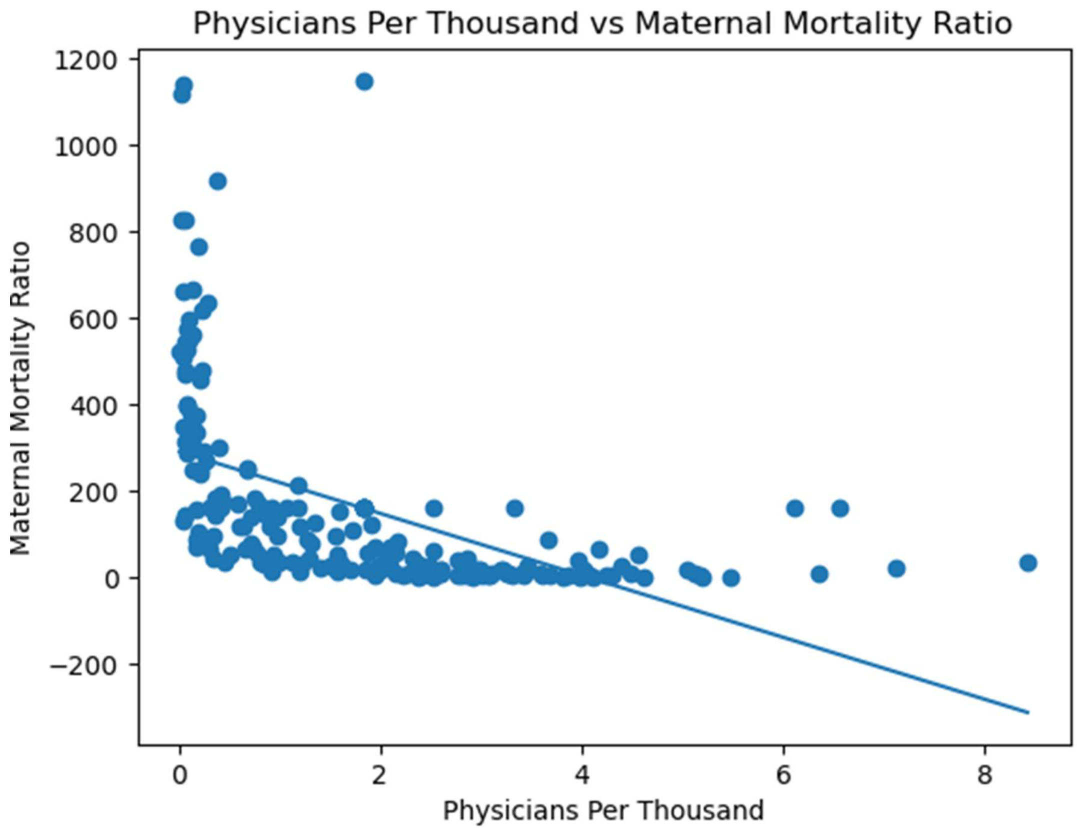
Life Expectancy vs Infant Mortality
Infant Mortality
Life Expectancy
0
20
40
60
80
55
60
65
70
75
80
85

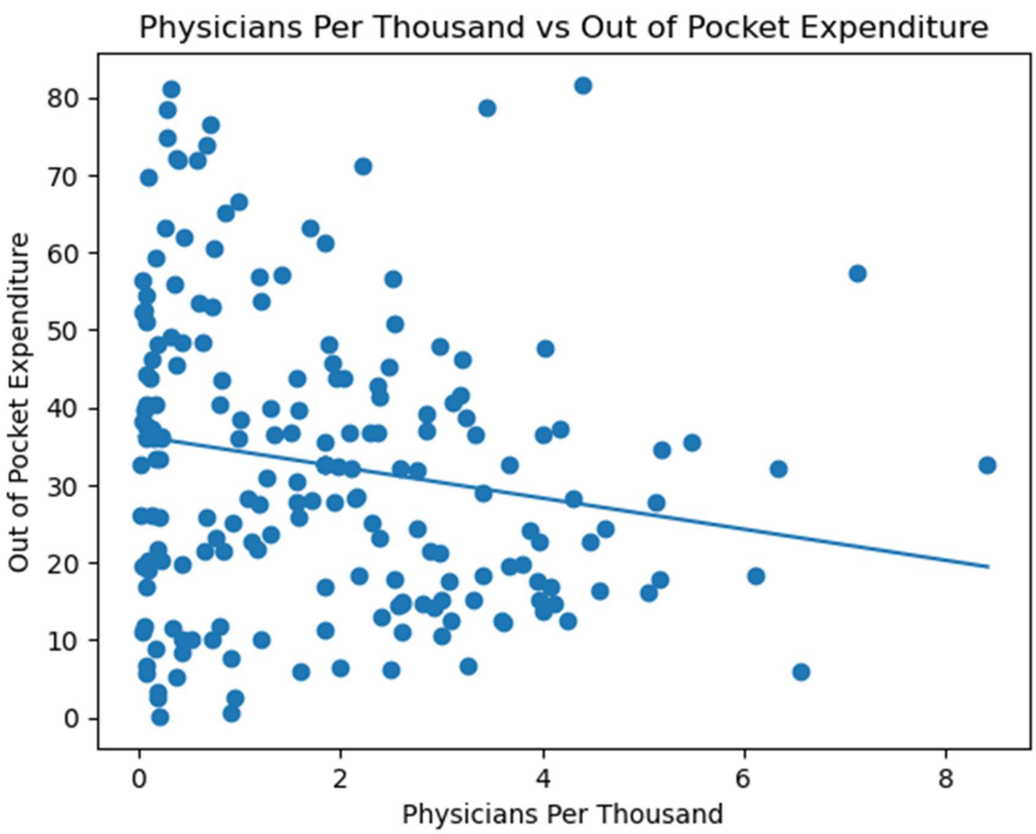

Life Expectancy vs Maternal Mortality Ratio
Maternal Mortality Ratio
Life Expectancy
1200
1000
800
600
400
200
0
−200
55
60
65
70
75
80
85

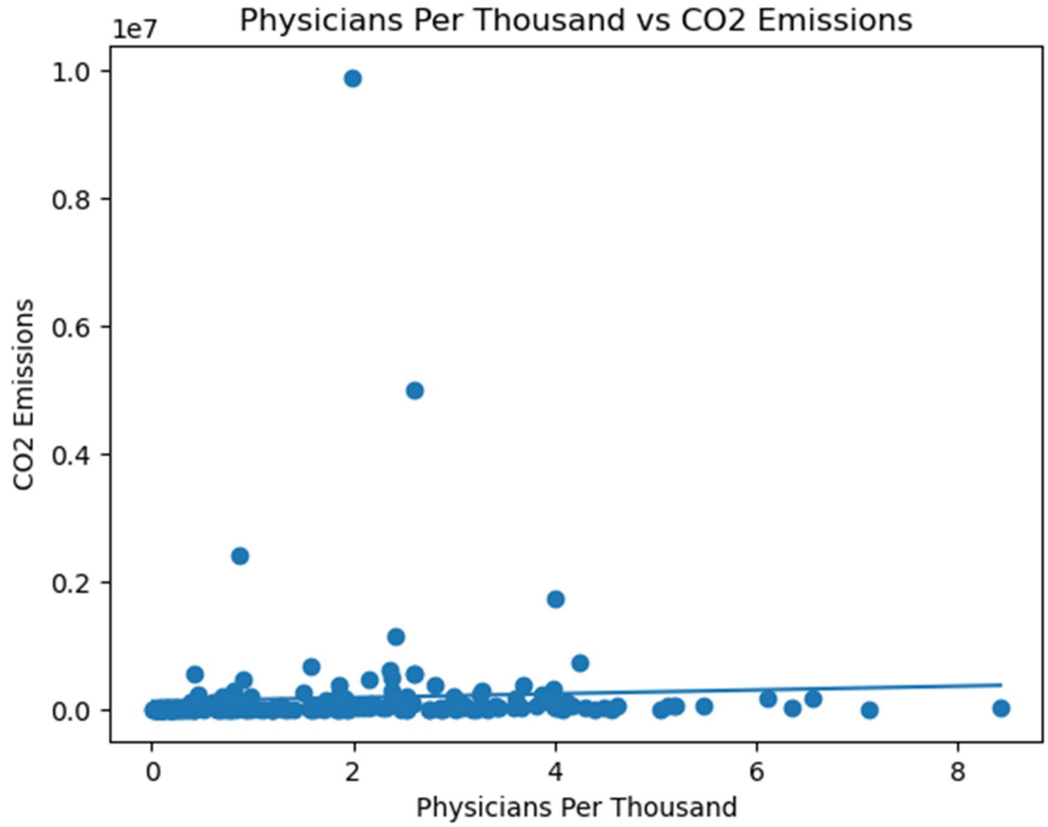

Life Expectancy vs Out of Pocket Expenditure
Out of Pocket Expenditure
Life Expectancy
80
70
60
50
40
30
20
10
0
55
60
65
70
75
80
85

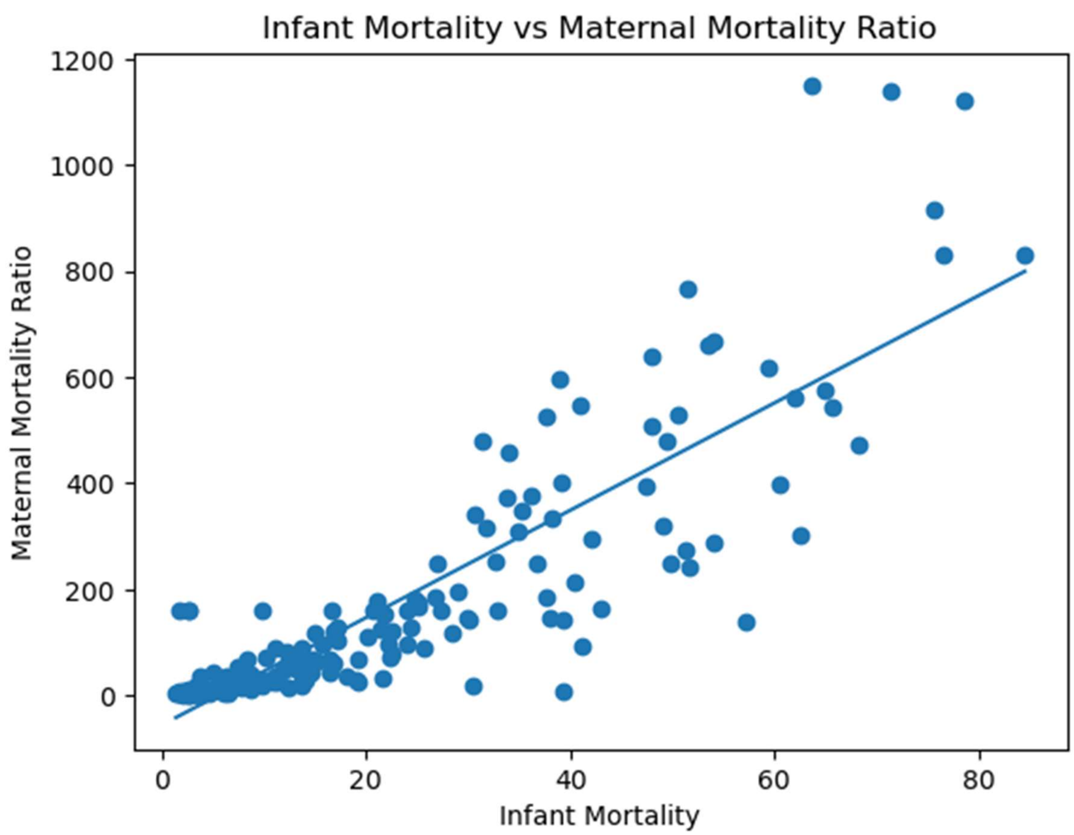

Life Expectancy vs CO2 Emissions
1e7
1.0
0.8
0.6
0.4
0.2
0.0
CO2 Emissions
55
60
65
70
75
80
85
Life Expectancy

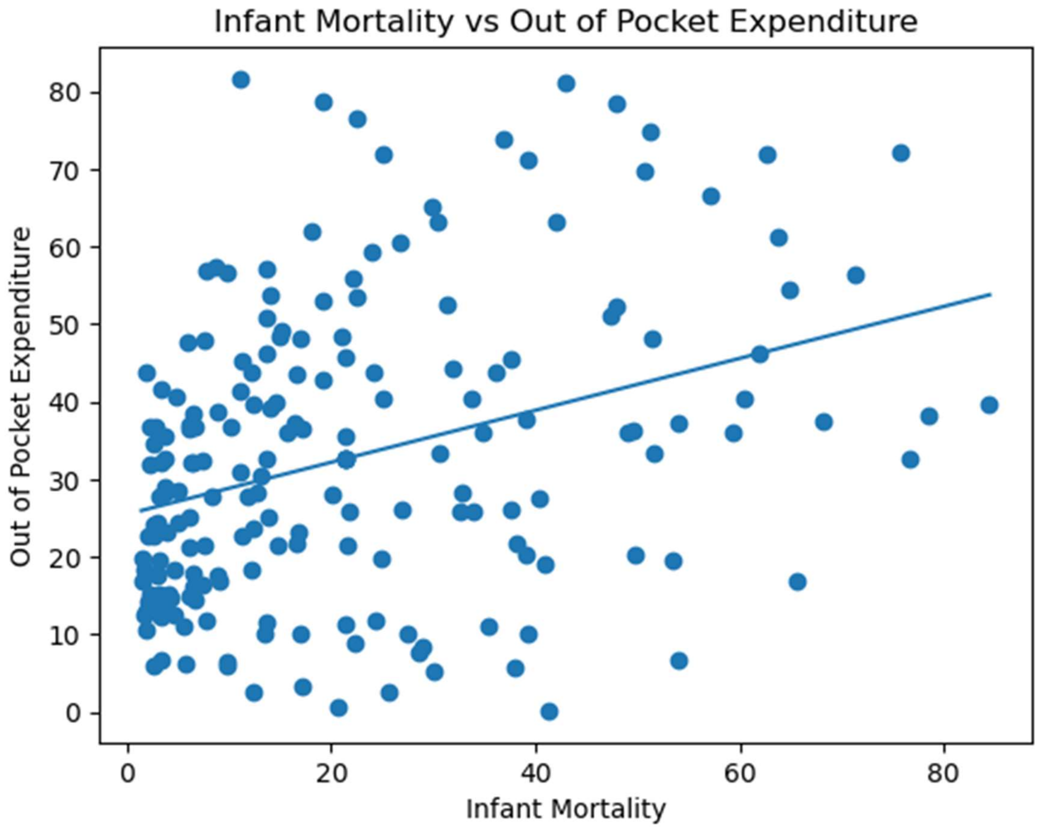

Physicians Per Thousand vs Infant Mortality
80
60
40
20
0
−20
Infant Mortality
0
2
4
6
8
Physicians Per Thousand

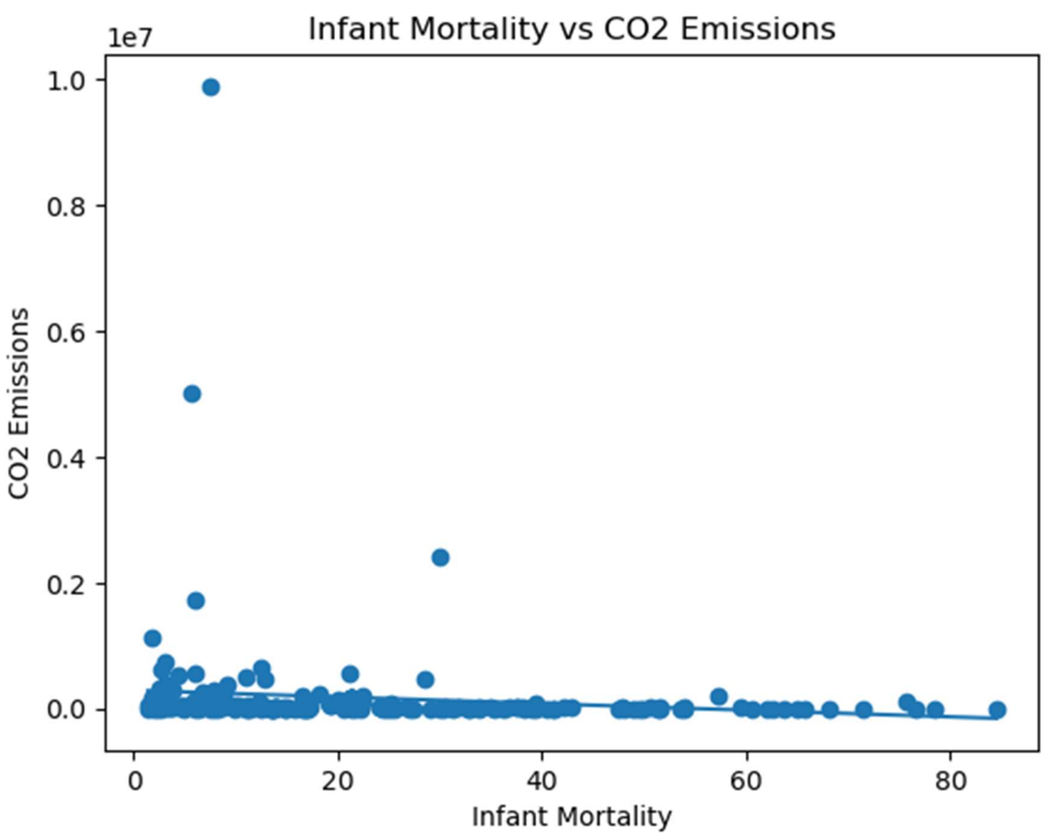

Physicians Per Thousand vs Maternal Mortality Ratio
Maternal Mortality Ratio
1200
1000
800
600
400
200
0
−200
0
2
4
6
8
Physicians Per Thousand

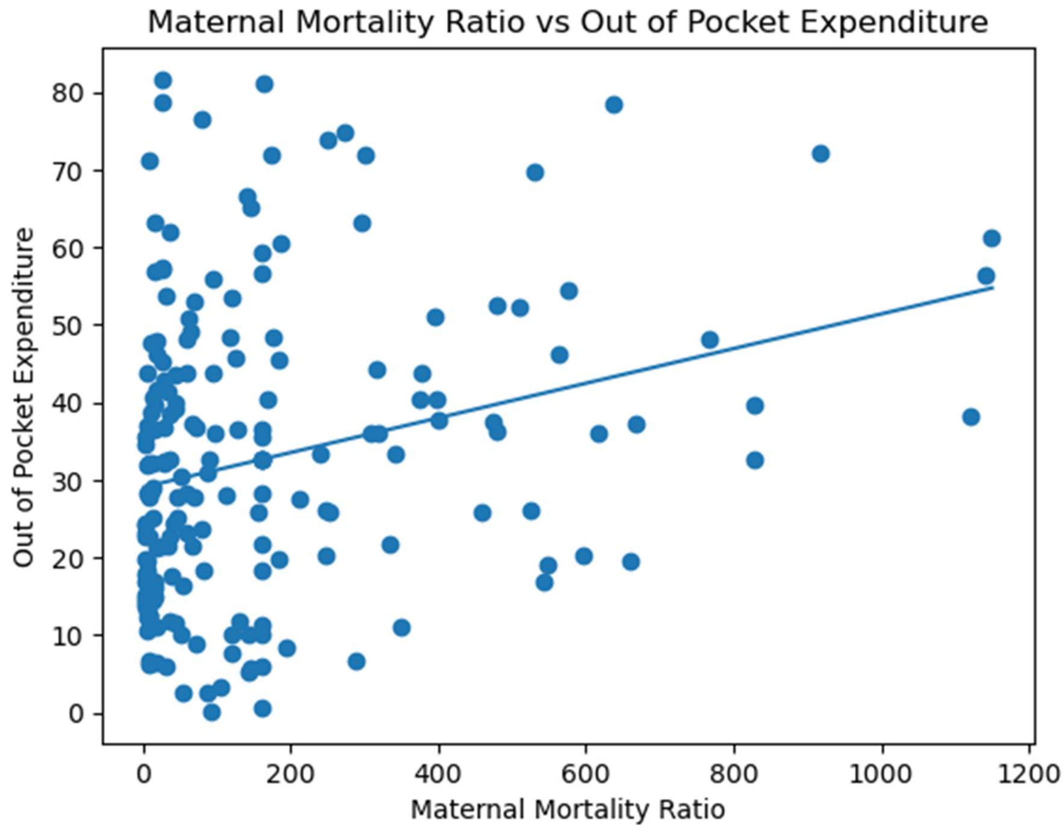

Physicians Per Thousand vs Out of Pocket Expenditure
Out of Pocket Expenditure
80
70
60
50
40
30
20
10
0
0
2
4
6
8
Physicians Per Thousand


Physicians Per Thousand vs CO2 Emissions
1e7
1.0
0.8
0.6
0.4
0.2
0.0
CO2 Emissions
0
2
4
6
8
Physicians Per Thousand

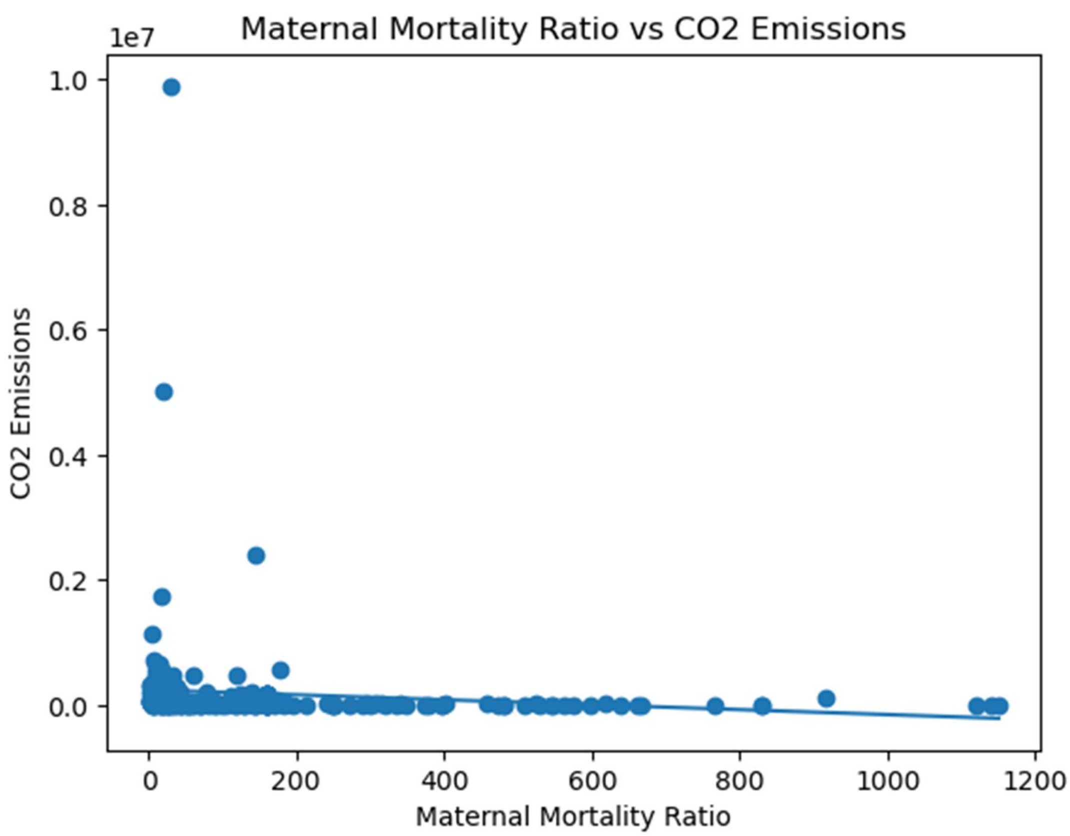

Infant Mortality vs Maternal Mortality Ratio
1200
1000
800
600
400
200
0
Maternal Mortality Ratio
0
20
40
60
80
Infant Mortality


Infant Mortality vs Out of Pocket Expenditure
Out of Pocket Expenditure
Infant Mortality
0
10
20
30
40
50
60
70
80

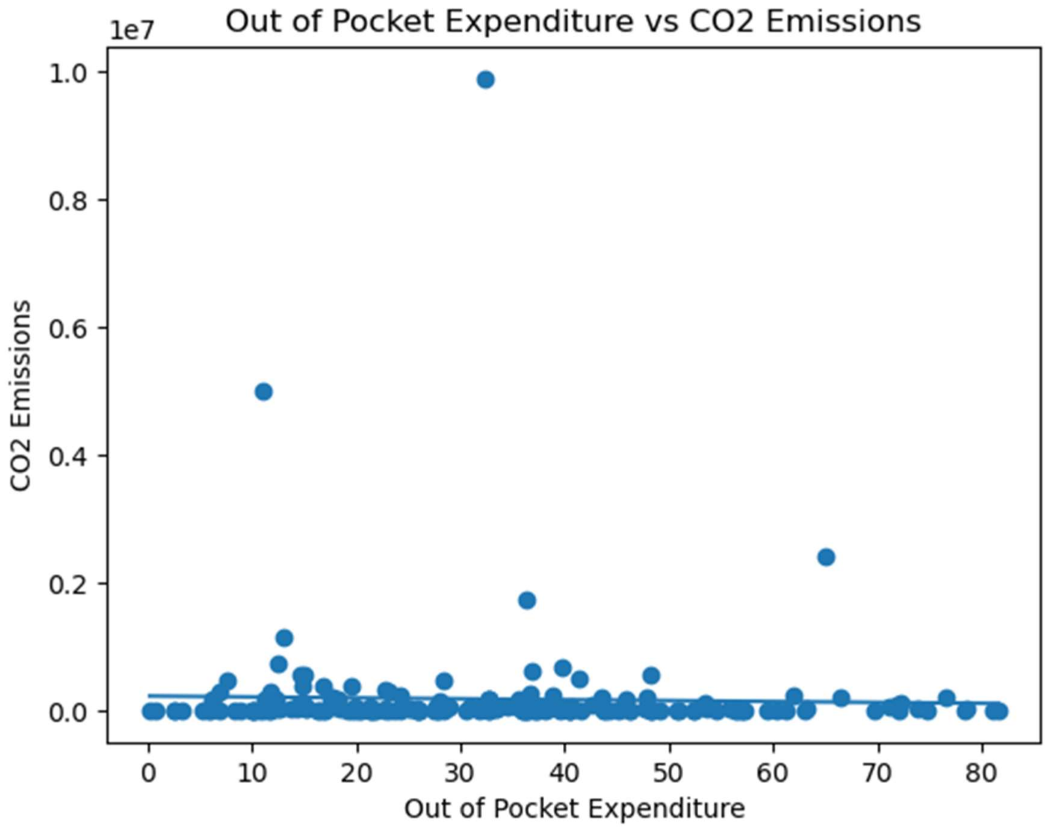

Infant Mortality vs CO2 Emissions
1e7
CO2 Emissions
Infant Mortality
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80



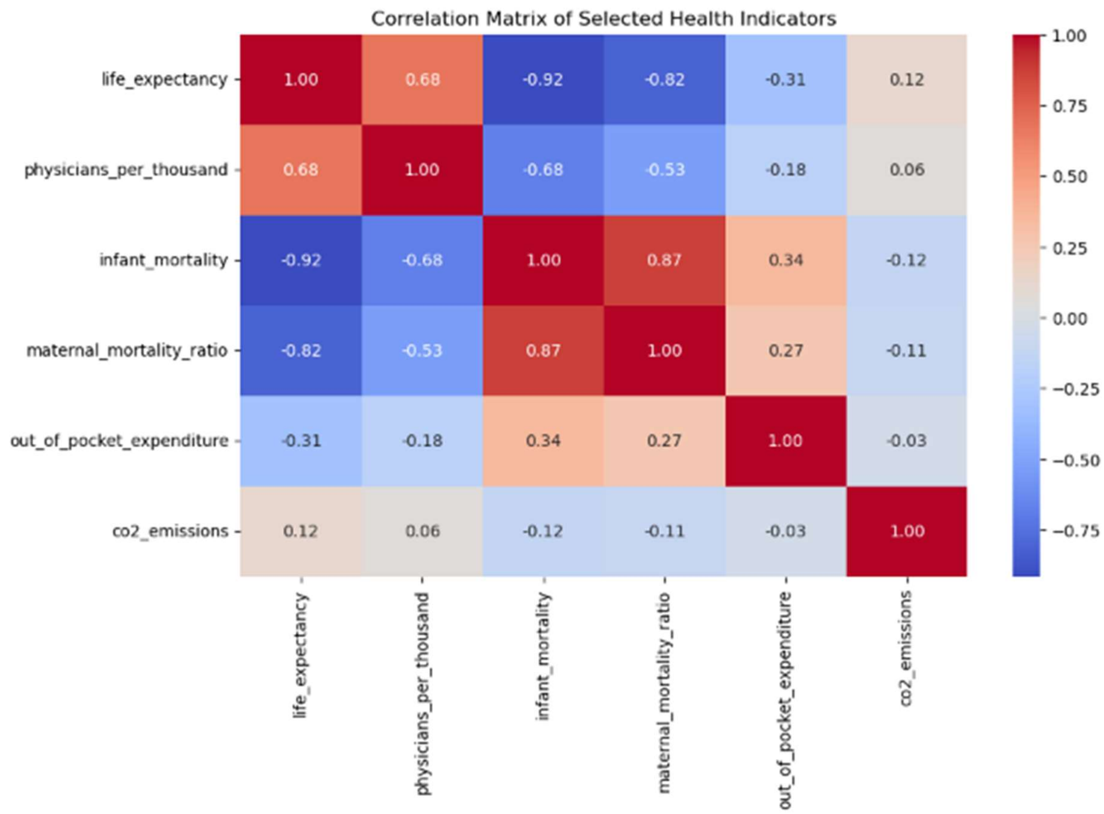
Maternal Mortality Ratio vs Out of Pocket Expenditure
Out of Pocket Expenditure
Maternal Mortality Ratio
0
10
20
30
40
50
60
70
80
0
200
400
600
800
1000
1200

Maternal Mortality Ratio vs CO2 Emissions
1e7
CO2 Emissions
Maternal Mortality Ratio
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
1200




To better understand the relationships between the variables in my dataset, I generated a correlation Matrix / Heatmap using Pearson correlation coefficients. This allowed me to explore how the selected health indicators interact with one another, and it provided crucial insights for designing a balanced and meaningful Country Health Index.



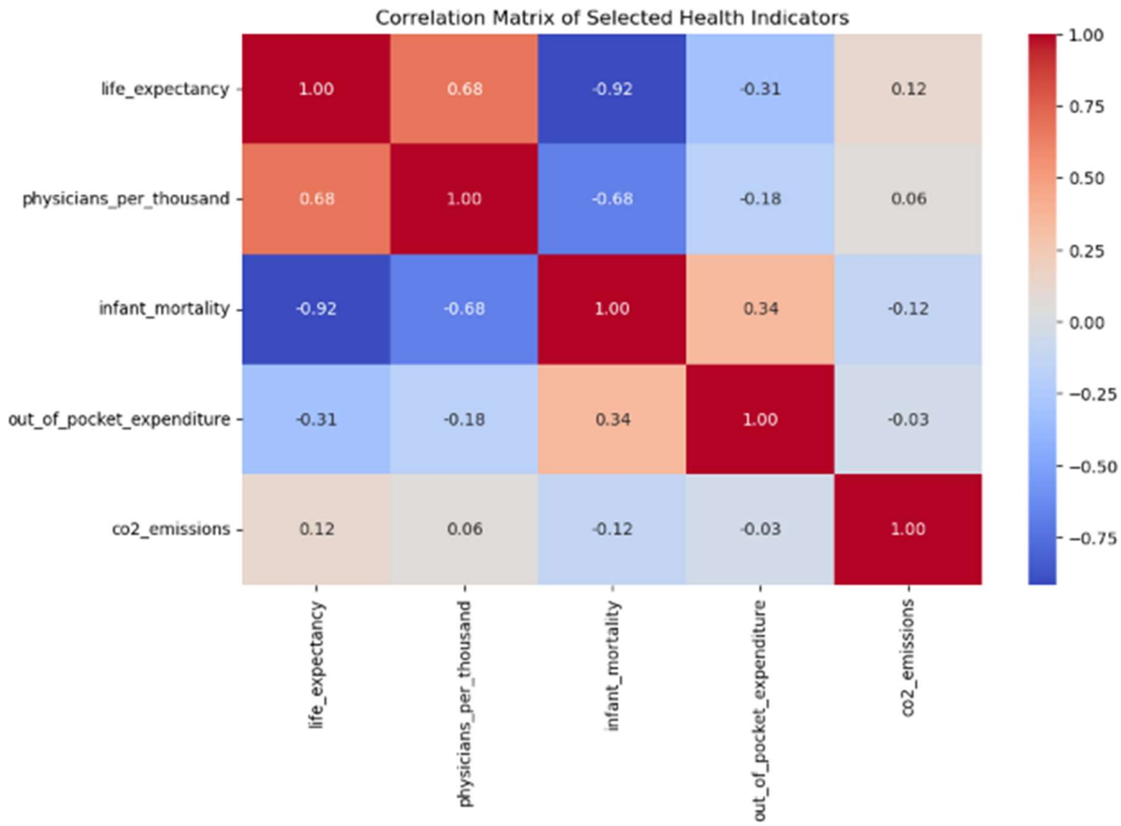

One of the notable findings is the **strong negative correlation** between life expectancy and infant mortality **(-0.92).** This is expected as countries with lower infant mortality rates generally tend to have higher average lifespans. Similarly, there is also a **strong negative correlation** between life expectancy and maternal mortality **(-0.82)**, reinforcing the idea that maternal and child health are fundamental to a country's overall health.

I also observed a **moderate positive correlation (0.68)** between life expectancy and physicians per thousand, suggesting that better access to healthcare professionals is associated with longer life spans. On the other hand, out-of-pocket expenditure has a **weak negative correlation (-0.31)** with life expectancy which may imply that in countries where patients need to bear more of the healthcare costs, life expectancies may lower.

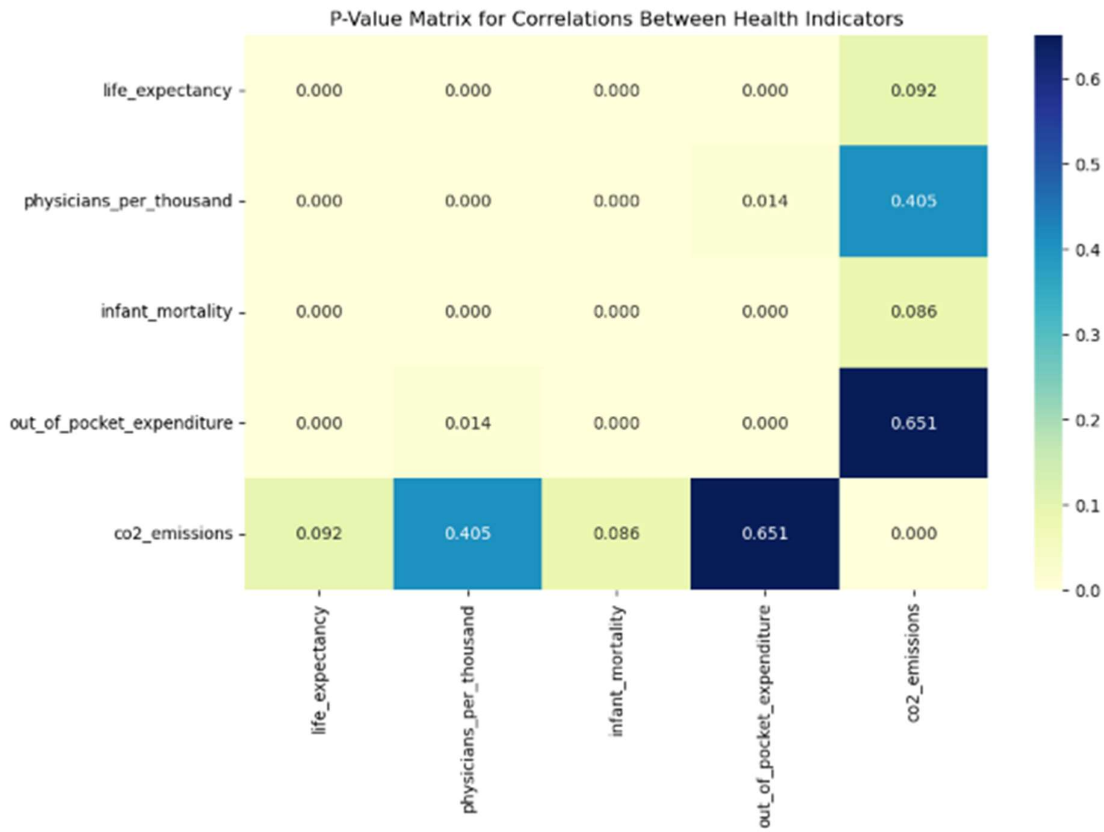
Surprisingly, CO2 emissions have very weak correlations between all other variables. While environmental factors are important, they may be much more complex than a direct relationship between CO2 emissions and health.

These correlations confirm that the variables that I have chosen are meaningfully related. I've decided that, to avoid redundancy, I have made the decision to drop the variable Maternal Mortality as it was highly correlated to infant mortality, showing multi-collinearity which I need to avoid, to have meaningful variables to construct the Country Health Index.

Below is the same Correlation Matrix with the Maternal Mortality Rate dropped from the dataset.



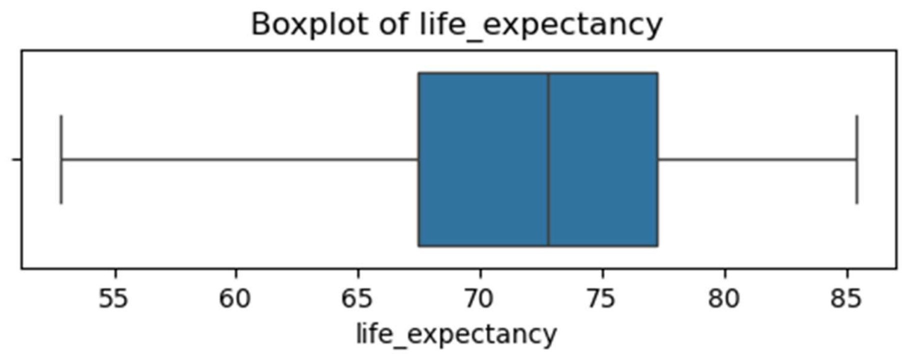

**Retrospective Note**: Upon completion of the index, I realised I should have done further analysis using P values to root out further multicollinearity, but I was too far into the development of the project. Here is a retrospective look at the P-Values using a Heatmap. To show what should have been addressed earlier but I failed to do so due to poor planning.

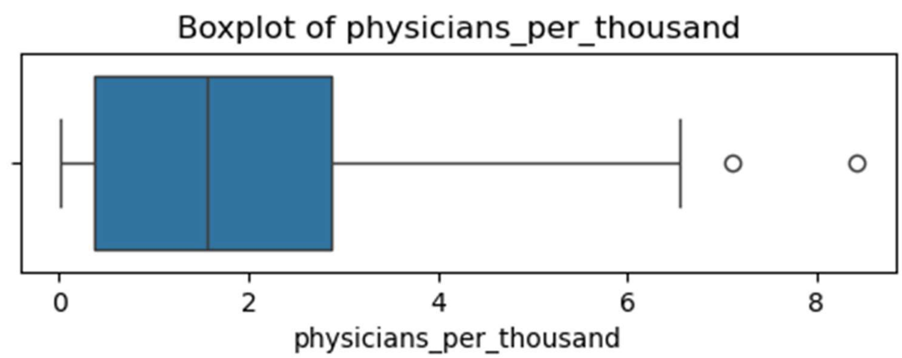


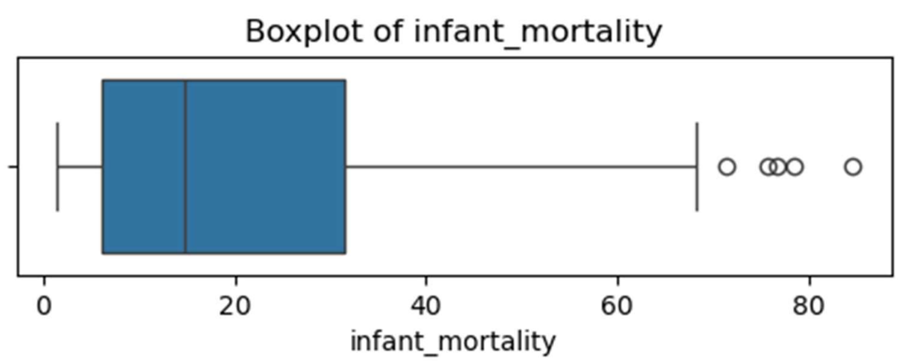

As you can see, a lot of the variables either have no P value or they have an extremely high P-Value meaning a lot of variables are not significant enough for use in the Index, BUT it is interesting to note that Out of Pocket Expenditure and Physicians per Thousand are highly significant with a **P-Value of 0.014** meaning they have a strong relationship.

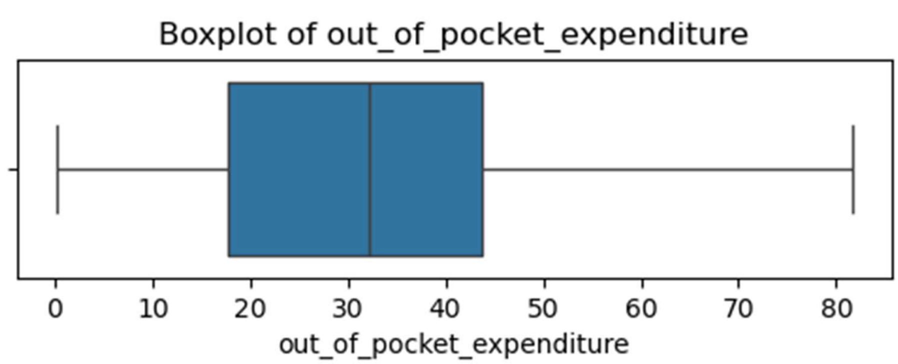
**Boxplots for Outlier Detection:**

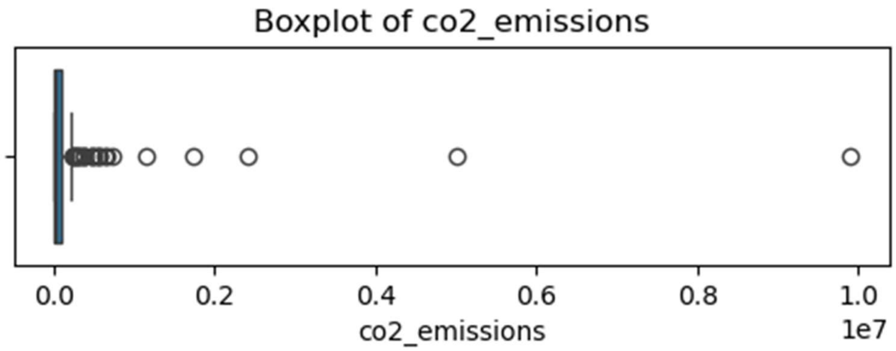
As part of my data exploration and multivariate analysis, I decided to user boxplots to check for outliers. Boxplots are especially useful as they highlight the median, inter-quartile range and the extreme values from a glance. By plotting each variable, I was able to see which countries had values that fell outside of the typical range, i.e. extremely high co2 emissions.


Boxplot of life_expectancy
55 60 65 70 75 80 85
life_expectancy


Boxplot of physicians_per_thousand
0 2 4 6 8
physicians_per_thousand


Boxplot of infant_mortality
0 20 40 60 80
infant_mortality


Boxplot of out_of_pocket_expenditure
0 10 20 30 40 50 60 70 80
out_of_pocket_expenditure




While outliers can skew averages and result in scaling difficulties if there is a disproportionate amount of them in the data. I decided to keep them in this case since they represent real and significant data points, for example, CO2 Emissions show extreme outliers, particularly for large, industrialised countries like the United States and China. The boxplots helped me make these decisions based on a clear understanding of each variable's spread and distribution.

## 5. Normalisation:

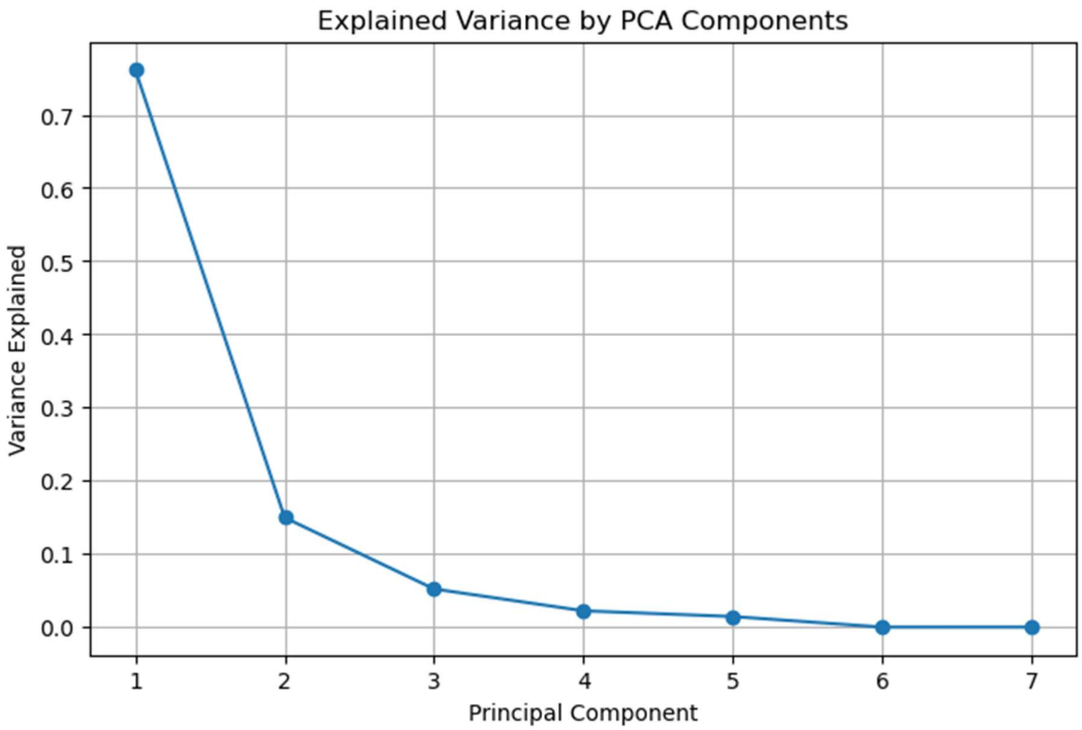
To ensure that all variables contributed fairly to the Country Health Index (CHI), I applied Min-Max normalisation to rescale the data to a common range between 0 and 1. This step is crucial as the original values were measured differently, using different ranges and different scales, for example life expectancy was typically in the 60-85 range while CO2 emissions ranges in the thousands. Without normalisation, larger values tend to dominate the Index calculations. Below is how the data appears after normalisation.

| | life_expectancy | physicians_per_thousand | infant_mortality | out_of_pocket_expenditure | co2_emissions | country_healthy_index | country_healthy_index_scaled | Country |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.358896 | 0.032105 | -0.559567 | -0.960688 | -0.000875 | 0.223983 | 0.223983 | Afghanistan |
| 1 | 0.788344 | 0.141498 | -0.077016 | -0.696560 | -0.000457 | 0.646229 | 0.646229 | Albania |
| 2 | 0.733129 | 0.203329 | -0.225030 | -0.342752 | -0.015162 | 0.655574 | 0.655574 | Algeria |
| 3 | 0.597536 | 0.394768 | -0.015644 | -0.444717 | -0.000046 | 0.716268 | 0.716268 | Andorra |
| 4 | 0.245399 | 0.023781 | -0.604091 | -0.407862 | -0.003506 | 0.271629 | 0.271629 | Angola |

Additionally, I applied negation to the negatively associated indicators such as out-of-pocket expenditure, co2 emissions and infant mortality, which need to go down rather than up to improve health outcomes in the target country.

## 6. Weighing and Aggregation

To objectively determine the contribution of each variable to the Country Health Index (CHI), I used Principal Component Analysis (PCA). PCA is a powerful technique for dimensionality reduction and simplification of large complex datasets into smaller sets while maintaining important patterns and trends between the variables. It transforms the original correlated variables into a set of uncorrelated components and orders them based on the amount of variance each component explains in the data set. This allowed me to identify the most significant patterns in the dataset.




I first applied PCA to the normalised data set. The explained variance plot showed revealed how much information each principal component captured. The first principal component (PC1) explained the largest proportion of the variance, indicating that it effectively summarised the combined health related variables of the countries into a single index.

```
PCA Loadings:
     life_expectancy  physicians_per_thousand  infant_mortality  \
PC1         0.587761                 0.425475          0.611484
PC2        -0.165839                -0.262353         -0.140935
PC3        -0.350187                 0.864637         -0.328404
PC4         0.054979                -0.047610          0.029922
PC5         0.708083                 0.016688         -0.705323

     out_of_pocket_expenditure  co2_emissions
PC1                   0.314391      -0.027468
PC2                   0.940059       0.009764
PC3                   0.129546       0.071695
PC4                  -0.009456       0.996858
PC5                   0.024005      -0.016857
```

After applying PCA to the dataset I decided to look at the PCA loadings which show how each variable shapes the index. The most influential indicators in defining the first principal component are:

- **Infant mortality (0.61)**
- **Life expectancy (0.59)**
- **Physicians per thousand (0.43)**
- **Out-of-pocket expenditure (0.31)**

These high loadings suggest that PC1 captures a composite view of health outcomes, access to care, life expectancy, physician density and financial burden. As a result of the extremely high variance explained by PC1, I have selected it as the basis for the Country Health Index (CHI). To make the index more intuitive, I decided to rescale the PC1 scores to a 0-100 scale where higher values represent healthier countries and lower scores represent less healthier countries.  Here you can see the top 10 healthiest countries and the top 10 least healthiest countries.

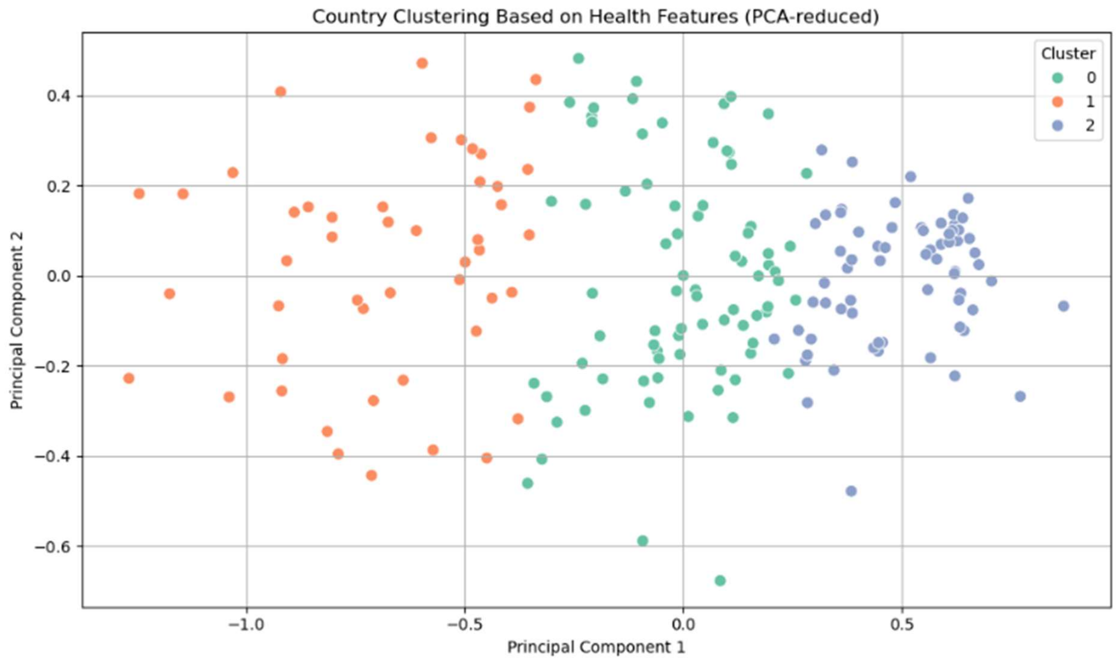
| | Country | country_healthy_index_scaled |
|---|---|---|
| 149 | San Marino | 100.000000 |
| 42 | Cuba | 95.357705 |
| 9 | Austria | 92.286942 |
| 76 | Iceland | 90.933661 |
| 168 | Sweden | 90.512733 |
| 82 | Israel | 90.289990 |
| 64 | Germany | 89.921472 |
| 60 | France | 89.790246 |
| 140 | Portugal | 89.337337 |
| 113 | Monaco | 89.192862 |

| | Country | country_healthy_index_scaled |
|---|---|---|
| 126 | Nigeria | 0.000000 |
| 33 | Central African Republic | 1.096999 |
| 34 | Chad | 4.348339 |
| 155 | Sierra Leone | 5.779753 |
| 53 | Equatorial Guinea | 10.713157 |
| 160 | Somalia | 11.108298 |
| 69 | Guinea | 16.011162 |
| 95 | Lesotho | 16.236137 |
| 31 | Cameroon | 16.358610 |
| 163 | South Sudan | 16.428377 |

This final scaled Country Health Index believe will offer a clear and interpretable measure of health performance per country.

## Clustering:

To further explore patterns between the different country's health characteristics, I employed a form of clustering known as K-Means Clustering after Principal Component Analysis to see how countries with similar health profiles are grouped. I decided to choose three clusters to see how they are displayed. Visualising these clusters using the first two principal components shows how diverse different health systems and outcomes are in different parts of the world.

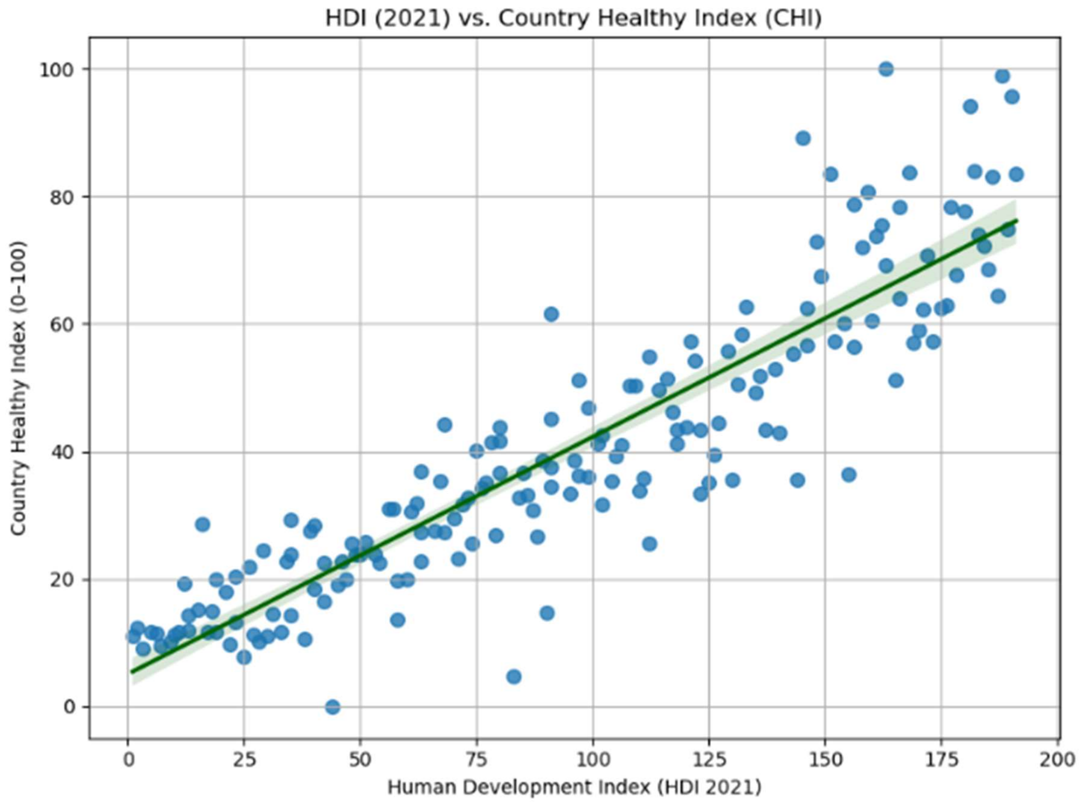



You can see here that the grouping of the countries dives perfectly in three, with the lowest ranking countries (countries with poorer health outcomes and poorer healthcare systems) to the left and the highest-ranking countries (countries with better health outcomes and better healthcare systems) on the right.

## 7. Links to other Indicators

To validate the Country Health Index (CHI) and better understand its alignment with broader global development measures, I compared it against the Human Development Index (HDI) for the year 2021. The HDI is a globally and universally recognised composite measure that incorporates life expectancy, education and income to assess a country's level of human development.

**Source:**
https://www.kaggle.com/datasets/iamsouravbanerjee/human-development-index-dataset/data




After merging the HDI dataset with my CHI data and applying min-max scaling to standardise the scores, I calculated the correlation between the two indices, which resulted in a very strong positive correlation (0.90). This high correlation suggests that countries performing well in terms of human development also tend to perform well in terms of health outcomes. The above regression plot shows a clear and consistent trend between the two indices, validating the structure and relevance of the CHI.

While CHI focuses solely on health specific indicators such as infant mortality, life expectancy, physicians per thousand etc., the strong correlation with HDI demonstrates that health is a key component of overall human development.

**Other Indicators Considered:**

While exploring other potential Indicators to compare with the CHI I came across the following Indicators that I wanted to compare:

- **WHO Universal Health Coverage (UHC) Index**
- **Global Health Security Index (GHS Index)**

However, I was unable to find adequate datasets for the above indices, at least any that were open sourced unfortunately.

### 8. Visualisation of the Results

To make the Country Health Index (CHI) results both interpretable and impactful, I employed a range of visualisations to highlight key insights and patterns across the dataset.

**Top 10 ranked countries statistics:**

## 1. San Marino

Life Expectancy: **85**

Physicians Per Thousand: **6.11**

Infant Mortality: **1.7**

Out of Pocket Expenditure: **18.3%**

## 2. Cuba

Life Expectancy: **78**

Physicians Per Thousand: **8.42**

Infant Mortality: **3.7**

Out of Pocket Expenditure: **32.6%**

## 3. Austria

Life Expectancy: **81**

Physicians Per Thousand: **5.17**

Infant Mortality: **2.9**

Out of Pocket Expenditure: **17.9%**

## 4. Iceland

Life Expectancy: **82**

Physicians Per Thousand: **4.08**

Infant Mortality: **1.5**

Out of Pocket Expenditure: **17.0%**

## 5. Sweden

Life Expectancy: **82**

Physicians Per Thousand: **3.98**

Infant Mortality: **2.2**

Out of Pocket Expenditure: **15.2%**

## 6. Israel

Life Expectancy: **82**

Physicians Per Thousand: **4.62**

Infant Mortality: **3.0**

Out of Pocket Expenditure: **24.4%**

## 7. Germany

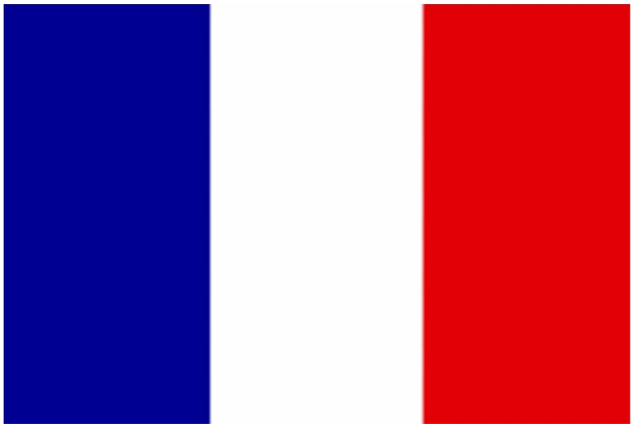
Life Expectancy: **80**

Physicians Per Thousand: **4.25**

Infant Mortality: **3.1**

Out of Pocket Expenditure: **12.5%**

## 8. France



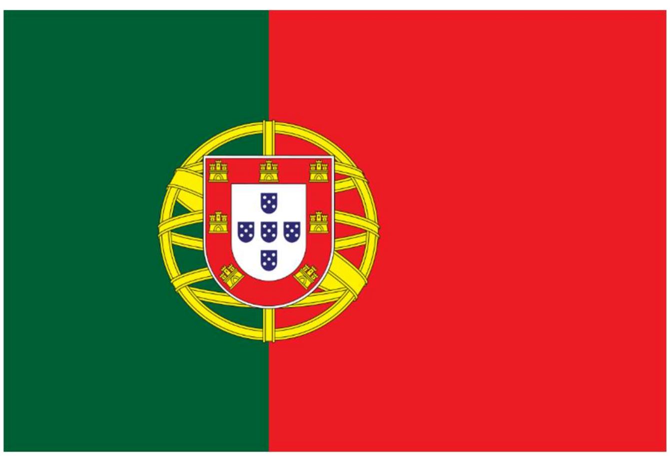
Life Expectancy: **82**

Physicians Per Thousand: **3.27**

Infant Mortality: **3.4**

Out of Pocket Expenditure: **6.8%**

## 9. Portugal



Life Expectancy: **81**

Physicians Per Thousand: **5.12**

Infant Mortality: **3.1**

Out of Pocket Expenditure: **27.7%**

## 10. Monaco

Life Expectancy: **72**

Physicians Per Thousand: **6.56**

Infant Mortality: **2.6**

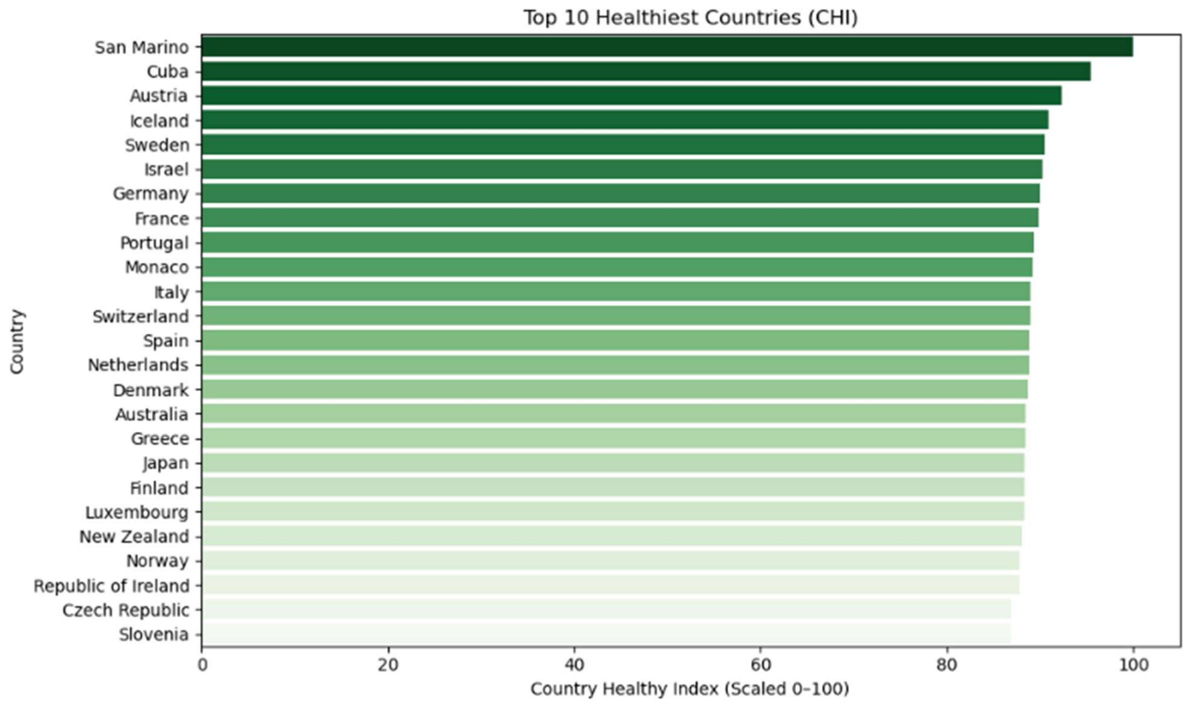
Out of Pocket Expenditure: **6.1%**

### Top 100 Healthy Countries:

| Rank | Country | Health Index (0-100) |
|---|---|---|
| 1 | San Marino | 100.000000 |
| 2 | Cuba | 95.357705 |
| 3 | Austria | 92.286942 |
| 4 | Iceland | 90.933661 |
| 5 | Sweden | 90.512733 |
| 6 | Israel | 90.289990 |
| 7 | Germany | 89.921472 |
| 8 | France | 89.790246 |
| 9 | Portugal | 89.337337 |
| 10 | Monaco | 89.192862 |

I first ranked all countries based on their CHI scores, sorted in descending order and extracted the top 100 healthiest countries. This list provides a simple yet concise ranked comparison between countries based on the health-related indicators used to create the index. The top performers tended to be high-income countries with strong healthcare systems, low infant mortality rates and widespread access to healthcare.

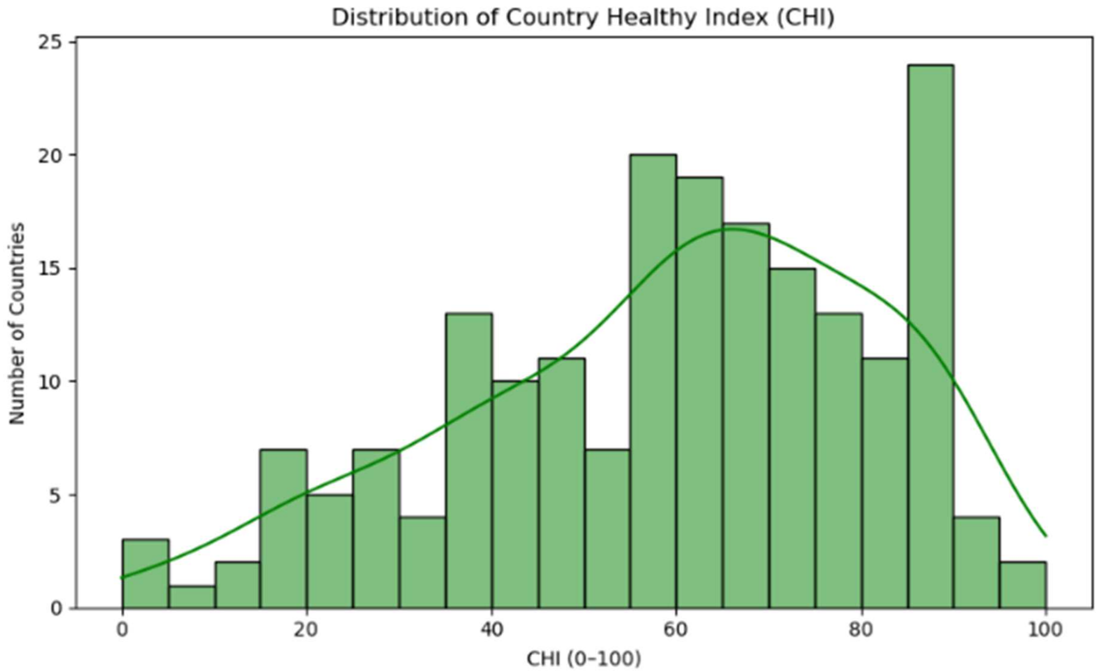
**Bar chart showing the top 25 healthiest countries:**




To further emphasis the leading countries, I created a horizontal bar chart of the Top 25 countries by CHI ranking. This chart offers a concise visual comparison of how closely ranked the top countries are and provides an intuitive sense of scale. Out of the top 25 countries with the highest Country Health Index rating, only 5 countries reside outside of Europe:

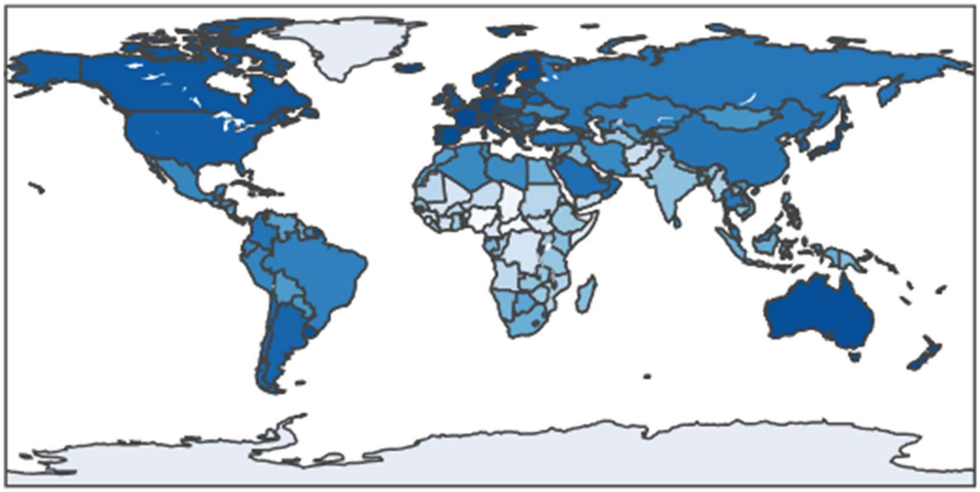
- 2. Cuba
- 6. Israel
- 16. Australia
- 18. Japan
- 21. New Zealand

**Histogram showing distribution of countries based of CHI ranking:**



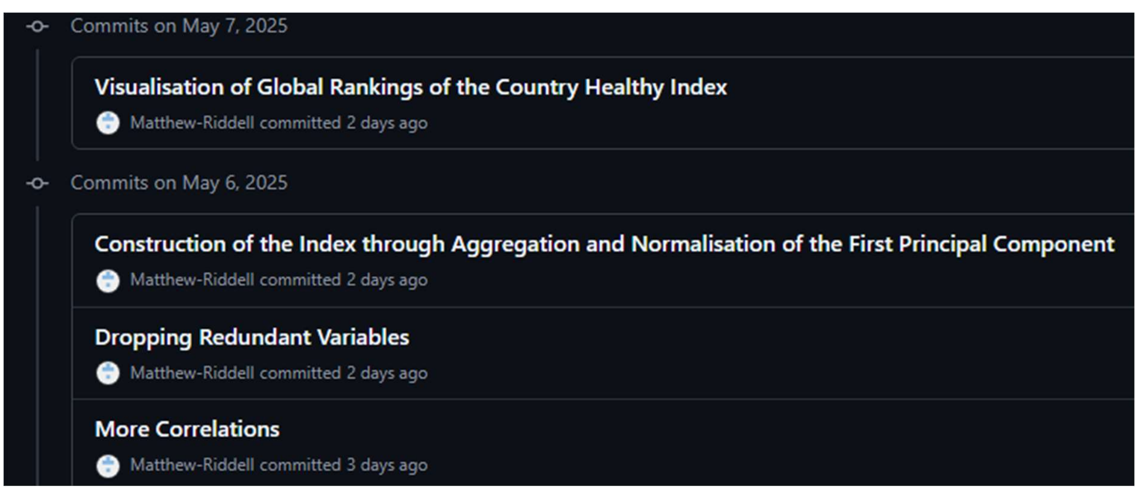

Next, I plotted a histogram showing the overall distribution of CHI scores across all countries. This visualisation revealed that while most countries were clustered in the mid-range (between 30-70) a lot of countries achieved exceptionally high scores in the 80-100 range. The line drawn on the graph helps to illustrate the skew towards the upper range of the Country Health Index.

**World Map: CHI by Country:**



Finally, I used Ploty Express to generate a choropleth map, which visually shows the CHI scores across each country on the world map. Countries are shaded in various intensities of blue based on their CHI scores. Darker shades mean higher CHI scores while lighter shades mean lower CHI scores. This map provides a geographically intuitive way to identify regional trends and disparities, that would normally go unnoticed when only viewing tables and graphs. For example, the high performing countries are generally grouped around Europe while lower performing countries are grouped around sub-Saharan Africa.

## 9. Version Control:



Throughout the development of the Country Health Index and the accompanying documentation, I used GitHub for version control. This enabled me to track changes, continuously progress on the project remotely from anywhere and document changes and manage reverts in case I need to go back to a previous version and back up changes and manage different stages of the project.

**GitHub Repo:** https://github.com/Matthew-Riddell/DAV_CA1

## 10. Conclusion:

Developing the Country Health Index allowed me to explore and quantify how different countries perform across a set of health-related indicators. Selecting well rounded indicators such as Life Expectancy and Physicians per Thousand and employing Min-Max normalisation across them and then employing Principal Component Analysis allowed me to develop a robust composite index that reflects overall the health of countries.

If I were able to develop this further, I would conduct further multivariate analysis and pull from more datasets to optimise and really hone down and get even better features for the final construction of the Index.

## 11. References:


World Health Organization (2025). *Health equity*. [online] World Health Organization. Available at: https://www.who.int/health-topics/health-equity#tab=tab_1.

ELGIRIYEWITHANA, N. (2023). *Global Country Information Dataset 2023*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023.

www.kaggle.com. (n.d.). *Human Development World Index*. [online] Available at: https://www.kaggle.com/datasets/iamsouravbanerjee/human-development-index-dataset/data.

WHO (2025). Universal health coverage (UHC). [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-(uhc).

GHS INDEX (2021). *The 2021 Global Health Security Index*. [online] GHS Index. Available at: https://ghsindex.org/.

VanderPlas, J. (2019). *In Depth: k-Means Clustering | Python Data Science Handbook*. [online] Github.io. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html.

**GitHub Repo for version control and management of the project:**

Matthew-Riddell (2025). *GitHub - Matthew-Riddell/DAV_CA1: CA1 for Data Analysis & Visualization*. [online] GitHub. Available at: https://github.com/Matthew-Riddell/DAV_CA1

**ChatGPT for assistance with the development of the CHI in Jupyter Notebook:**

ChatGPT. (2025). *ChatGPT - New chat*. [online] Available at: https://chatgpt.com/share/68166c24-14cc-8007-a343-b1b7946badd2.