



# Visualization of Health Data

Veronica Castro Alvarez and Ching-yu Huang<sup>(✉)</sup>

School of Computer Science, 1000 Morris Avenue, Union, NJ 07083, USA  
{Castroav, chuang}@kean.edu

**Abstract.** As data becomes more accessible, visualization methods are needed to help make sense of the information. Analyzing and visualizing data helps the public to better recognize the patterns and connections between different data-sets. By using visual elements such as graphs, charts, and maps, it is easier to see and understand the trends and outliers in data. This project aims to study the correlation between environmental factors and public health. Large sets of data pertaining to the environment and health were gathered from open data sources. The tool used to analyze and visualize the collected data is Tableau, which is a software program that is used to transform data into dashboards and visuals such as treemaps, histograms, or area charts. For this project, the data will be displayed through charts and interactive maps that will be created through this software.

**Keywords:** Environment · Health · Tableau · Visualization

## 1 Introduction

The environment can greatly impact individuals' health, and one of the biggest environmental factors that contributes to disease is pollution. There have been various studies and articles published on the effects of environmental factors on health and disease. According to an article published in *The Lancet*, a peer-reviewed medical journal, "Pollution is the largest environmental cause of disease and premature death in the world today. Diseases caused by pollution were responsible for an estimated 9 million premature deaths in 2015—16% of all deaths worldwide" [1]. Some of the most prominent and unresolved issues related to public health are air pollution, drinking water, soil pollution, and weather conditions [2]. Outdoor air pollution seems to be negatively affecting health outcomes at increasing rates, with respiratory diseases, such as asthma, and mortality being the most common and prevalent health outcomes [3].

These pollution factors, in addition to inadequate water sanitation, agricultural practices, built environments, and climate change all play a major role in the likelihood of developing a disease, and people are exposed to these risk factors every day in their homes and communities. There is also a need for safe water around the globe, and research has been published on the different areas of water and health, such as socioeconomy of water, water quality, water treatment, water microbiology, water sanitation, and water resources [4]. For example, study conducted in 26 Sub-Saharan African countries found that the "time spent walking to a household's main water source was found to be a significant determinant of under-five child health" [5].

Another important aspect to the burden of disease and death is income level, as individuals with low economic standing and those in developing countries usually have poor water quality and access [6]. “Nearly 92% of pollution-related deaths occur in low-income and middle-income countries and, in countries at every income level, disease caused by pollution is most prevalent among minorities and the marginalised” [1]. Additionally, individuals living at a low-income level are more at risk of death from disease in countries where they do not have any access to or have limited access to healthcare. For example, a recent study on the gradient between child’s health and family income in Canada did not find that children from low-income families suffered more from poor health than children from high-income families (contrary to previous studies from the United States). However, “the contrast between Canadian and U.S. children may reflect the effects of universal health insurance in Canada” [7].

When comparing low-income and middle-income countries, both power equality and per capita income seem to be important factors in relation to health performance, “while only power equality appears to be a factor in explaining population health in middle-income countries” [8]. However, neither power equality nor per capita income seem to be a factor in high-income countries, but rather “expenditures on health services relative to GDP and sanitation access” appear to be factors in these countries [8].

Although there have been research studies published on the areas of public health and the environment, there have not been many studies which use visualization methods to display the data associated with both of these areas. Big data visualization can be used to communicate large amounts of information that otherwise may not be as easy to understand by a general audience without a more in-depth look. This study aims to expand upon previous research of this topic by visualizing the available open data in order to make the information easier to understand for a general public audience.

## 2 Methods

Data was retrieved from two online open data sources, the World Health Organization’s Global Health Observatory and World Bank Open Data. The datasets pertaining to health factors were retrieved from the Global Health Observatory, while the datasets pertaining to environmental factors were retrieved from World Bank Open Data. The original formats of the datasets were either Comma-Separated Values (CSV) files or Microsoft Excel files.

### 2.1 Interpreting the Data

After gathering all of the datasets, each dataset was interpreted using Tableau’s Data Interpreter, which is a feature that can be used when connecting to Excel files to clean and transform the data. Data that is provided in Excel files are usually made to be easy to read with the human interface in mind, so they may often include titles, stacked headers, or empty rows and columns to make the file more visually appealing and easy to understand. This was the case for many of the datasets used in this study. However, these visually appealing aspects of the file can make the data more difficult for Tableau to interpret. Tableau provides this data interpretation feature to help identify the

structure of the data in the Excel file and detect if there are any things such as titles, stacked headers, or empty cells in the file. Once any of these aspects have been detected, Tableau's data interpreter converts the data into the proper format to be used for analysis. One of the datasets retrieved from World Bank Open Data was of Total Population by Country per Year. Figure 1 shows how this dataset appeared in Tableau before using Tableau's Data Interpreter to interpret the data, and Fig. 2 shows how the data appeared after using the Data Interpreter.

Sort fields Data source order ▾ <input type="checkbox"/> Show aliases <input type="checkbox"/> Show hidden fields 267 rows								
Abc	Abc	Abc	Abc	#	#	#	#	#
Data	Data	Data	Data	Data	Data	Data	Data	Data
F1	F2	F3	F4	F5	F6	F7	F8	F9
Data Source	World Developm...	null	null	null	null	null	null	
Last Updated Date	3/21/2019	null	null	null	null	null	null	
Country Name	Country Code	Indicator Name	Indicator Code	1,960	1,961	1,962	1,963	
Aruba	ABW	Population, total	SP.POP.TOTL	54,211	55,438	56,225	56,695	
Afghanistan	AFG	Population, total	SP.POP.TOTL	8,996,351	9,166,764	9,345,868	9,533,954	
Angola	AGO	Population, total	SP.POP.TOTL	5,643,182	5,753,024	5,866,061	5,980,417	
Albania	ALB	Population, total	SP.POP.TOTL	1,608,800	1,659,800	1,711,319	1,762,621	
Andorra	AND	Population, total	SP.POP.TOTL	13,411	14,375	15,370	16,412	
Arab World	ARB	Population, total	SP.POP.TOTL	92,490,932	95,044,497	97,682,294	100,411,076	
United Arab Emir...	ARE	Population, total	SP.POP.TOTL	92,634	101,078	112,472	125,566	
Argentina	ARG	Population, total	SP.POP.TOTL	20,619,075	20,953,077	21,287,682	21,621,840	

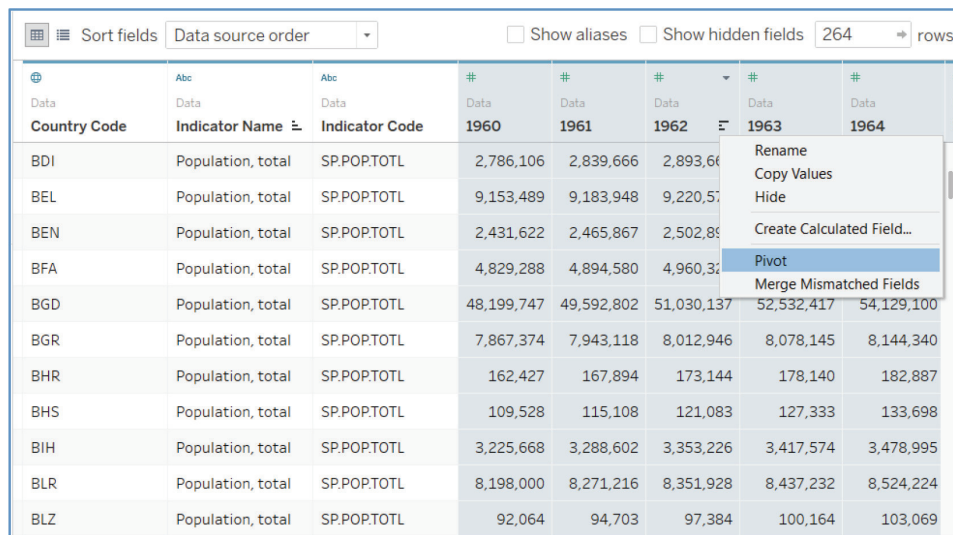
Fig. 1. Dataset before data interpretation

Sort fields Data source order ▾ <input type="checkbox"/> Show aliases <input type="checkbox"/> Show hidden fields 264 rows							
⊕	⊕	Abc	Abc	#	#	#	#
Data	Data	Data	Data	Data	Data	Data	Data
Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963
Aruba	ABW	Population, total	SP.POP.TOTL	54,211	55,438	56,225	5
Afghanistan	AFG	Population, total	SP.POP.TOTL	8,996,351	9,166,764	9,345,868	9,53
Angola	AGO	Population, total	SP.POP.TOTL	5,643,182	5,753,024	5,866,061	5,98
Albania	ALB	Population, total	SP.POP.TOTL	1,608,800	1,659,800	1,711,319	1,76
Andorra	AND	Population, total	SP.POP.TOTL	13,411	14,375	15,370	1
Arab World	ARB	Population, total	SP.POP.TOTL	92,490,932	95,044,497	97,682,294	100,41
United Arab Emir...	ARE	Population, total	SP.POP.TOTL	92,634	101,078	112,472	12
Argentina	ARG	Population, total	SP.POP.TOTL	20,619,075	20,953,077	21,287,682	21,62
Armenia	ARM	Population, total	SP.POP.TOTL	1,874,120	1,941,491	2,009,526	2,07
American Samoa	ASM	Population, total	SP.POP.TOTL	20,013	20,486	21,117	2
Antigua and Bar...	ATG	Population, total	SP.POP.TOTL	55,339	56,144	57,144	5

Fig. 2. Dataset after data interpretation

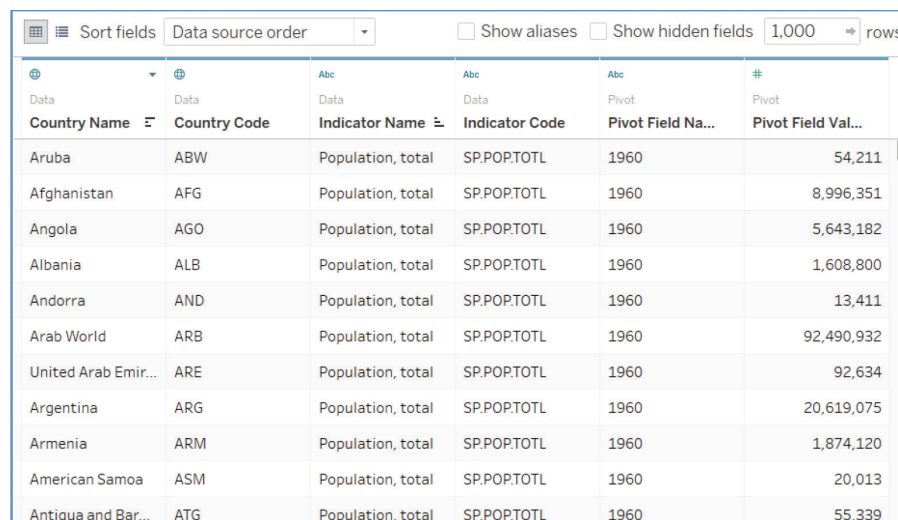
## 2.2 Pivoting the Data

After interpreting the datasets, they then had to be transposed from crosstab format (wide format) to columnar format (long format). Most of the datasets that were retrieved were originally in crosstab format, where there was one separate column for each year and each year was a column header. However, the ideal structure for a data table that is to be used for analysis is columnar format, where each row represents an observation belonging to a particular category, such as Year or Number of Cases. To transpose the tables from crosstab format to columnar format, this was done using Tableau's Pivot function, as shown in Figs. 3 and 4.



Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964
BDI	Population, total	SP.POP.TOTL	2,786,106	2,839,666	2,893,666		
BEL	Population, total	SP.POP.TOTL	9,153,489	9,183,948	9,220,517		
BEN	Population, total	SP.POP.TOTL	2,431,622	2,465,867	2,502,895		
BFA	Population, total	SP.POP.TOTL	4,829,288	4,894,580	4,960,347		
BGD	Population, total	SP.POP.TOTL	48,199,747	49,592,802	51,030,137	52,532,417	54,129,100
BGR	Population, total	SP.POP.TOTL	7,867,374	7,943,118	8,012,946	8,078,145	8,144,340
BHR	Population, total	SP.POP.TOTL	162,427	167,894	173,144	178,140	182,887
BHS	Population, total	SP.POP.TOTL	109,528	115,108	121,083	127,333	133,698
BIH	Population, total	SP.POP.TOTL	3,225,668	3,288,602	3,353,226	3,417,574	3,478,995
BLR	Population, total	SP.POP.TOTL	8,198,000	8,271,216	8,351,928	8,437,232	8,524,224
BLZ	Population, total	SP.POP.TOTL	92,064	94,703	97,384	100,164	103,069

Fig. 3. Dataset before data pivot



Country Name	Country Code	Indicator Name	Indicator Code	Pivot Field Name	Pivot Field Value
Aruba	ABW	Population, total	SP.POP.TOTL	1960	54,211
Afghanistan	AFG	Population, total	SP.POP.TOTL	1960	8,996,351
Angola	AGO	Population, total	SP.POP.TOTL	1960	5,643,182
Albania	ALB	Population, total	SP.POP.TOTL	1960	1,608,800
Andorra	AND	Population, total	SP.POP.TOTL	1960	13,411
Arab World	ARB	Population, total	SP.POP.TOTL	1960	92,490,932
United Arab Emir...	ARE	Population, total	SP.POP.TOTL	1960	92,634
Argentina	ARG	Population, total	SP.POP.TOTL	1960	20,619,075
Armenia	ARM	Population, total	SP.POP.TOTL	1960	1,874,120
American Samoa	ASM	Population, total	SP.POP.TOTL	1960	20,013
Antigua and Bar...	ATG	Population, total	SP.POP.TOTL	1960	55,339

Fig. 4. Dataset after data pivot

### 2.3 Joining the Data

After interpreting, pivoting, and filtering the datasets, they were then joined together by the fields Country Name and Year by left join using Tableau's Join function. This function allows sperate tables that are related by specific fields, such as Country and Year, to be combined on those common fields. This joining results in a virtual table that can then be used for analyzation and visualization. Figure 5 shows how the tables containing data on Cholera, Diphtheria, Malaria, Air Pollution, Arable Land, Renewable Internal Freshwater Resources, and Crude Death Rate were joined to the table containing on Country Information.

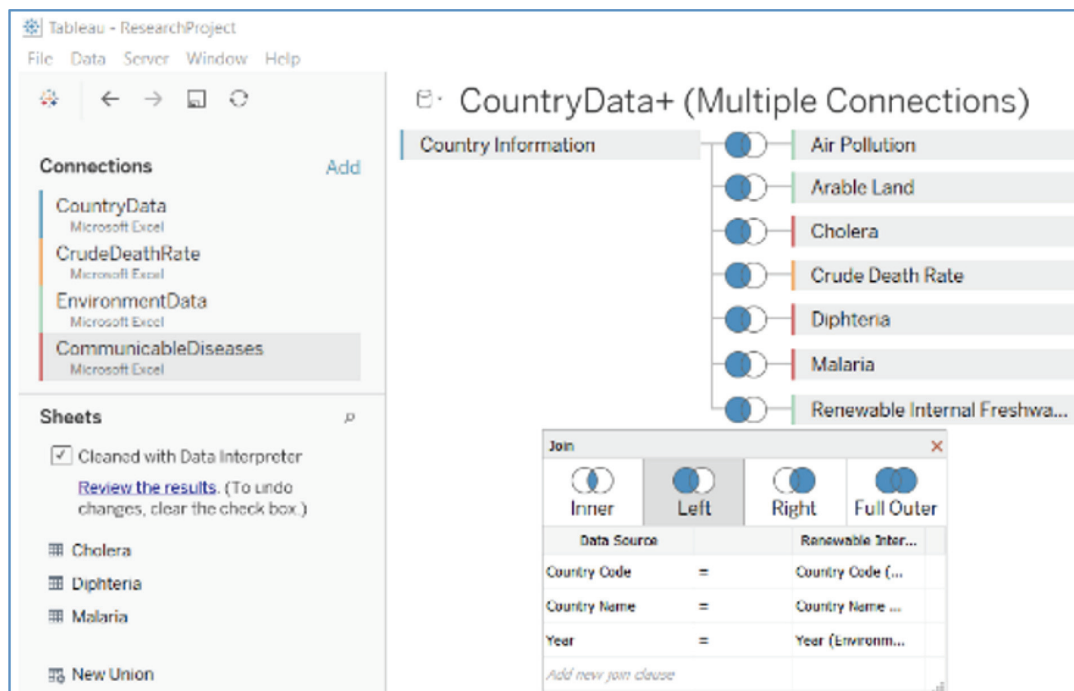
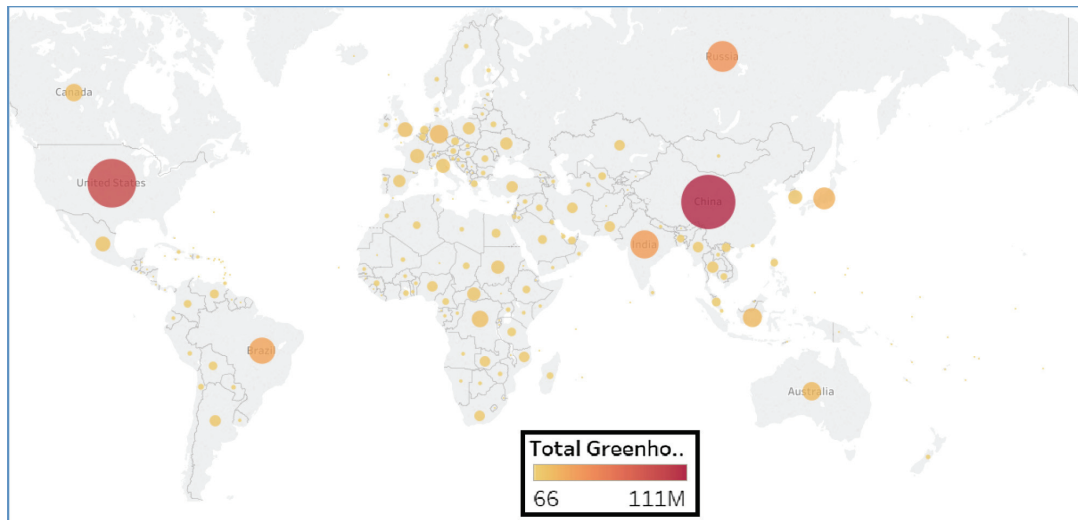


Fig. 5. Tableau's Join function used to join datasets

## 3 Experiment Results

Although the final data set consists of 23 columns, only 13 of those fields were used for this project. The environmental factors examined were: Total Greenhouse Gas Emissions, Carbon Dioxide (CO<sub>2</sub>) Emissions, Methane Emissions, Nitrous Oxide Emissions, and Renewable Internal Freshwater Emissions Per Capita. The health factors examined, which focused on communicable diseases, were: Number of Reported Cases of Cholera, Number of Reported Cases of Diphtheria, and Number of Reported Cases of Malaria.

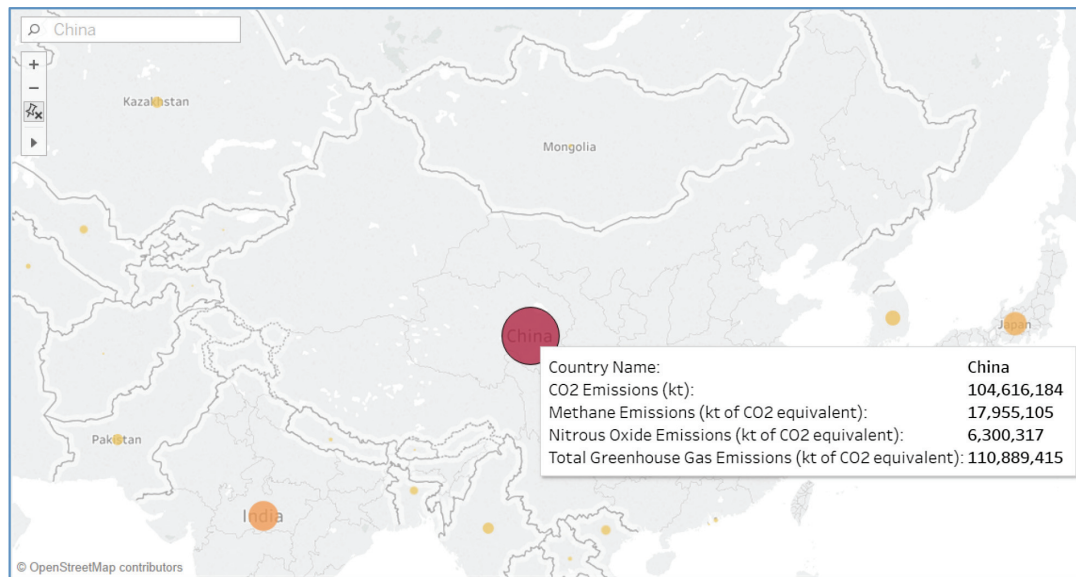
The first factors that were examined for this project were the environmental factors which were gathered from World Bank Open Data [9]. Figure 6 shows the Total Greenhouse Gas Emissions (in units of kt of CO<sub>2</sub> equivalent) by Country for the years 2000–2012. Some of the greenhouse gases that are summed together to determine this variable are Carbon Dioxide (CO<sub>2</sub>) emissions (in units of kt), Methane emissions (in units of kt of CO<sub>2</sub> equivalent), and Nitrous Oxide emissions (in units of kt of CO<sub>2</sub> equivalent). The following symbol map is one of the numerous charts that Tableau offers to visualize data. From the symbol map, it is shown that China emits the most greenhouse gases than any other country, followed by the United States of America.



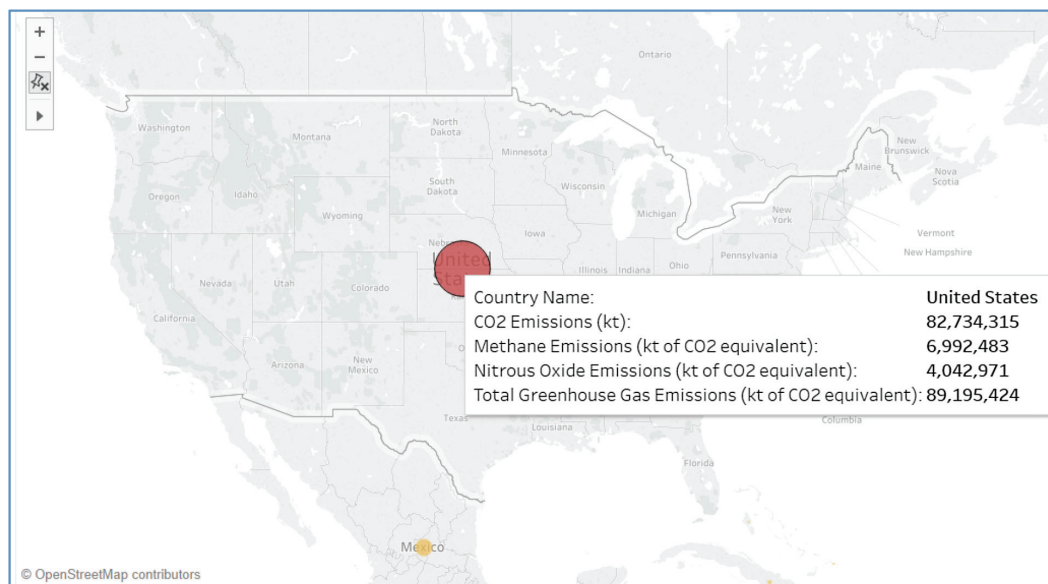
**Fig. 6.** Total Greenhouse Gas Emissions by Country

Tableau allows the user to zoom into specific areas of the map in order to emphasize any prominent countries or areas. Figure 7 shows the map zoomed into China, and Fig. 8 shows the map zoomed into the United States, the two countries with the highest greenhouse gas emissions. When the user hovers over the symbol, a tooltip appears to show the details for that selected country. The user can select which variables are to appear in the tooltip. In the tooltip used for this symbol map, the Country Name, CO<sub>2</sub> Emissions, Methane Emissions, Nitrous Oxide Emissions, and Total Greenhouse Gas Emissions are shown for each country.



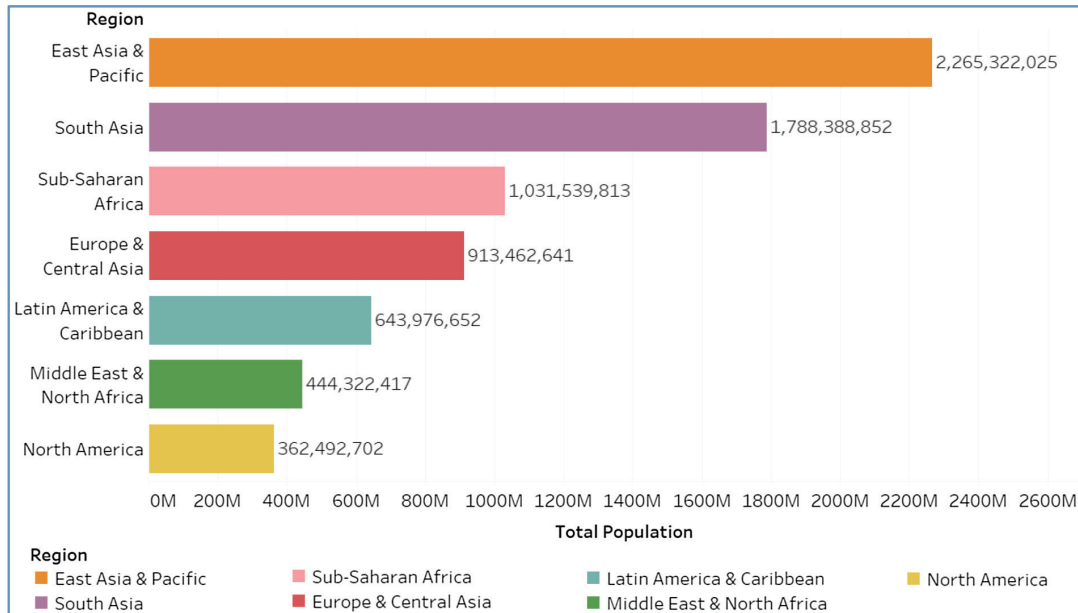


**Fig. 7.** China's Total Greenhouse Gas Emissions



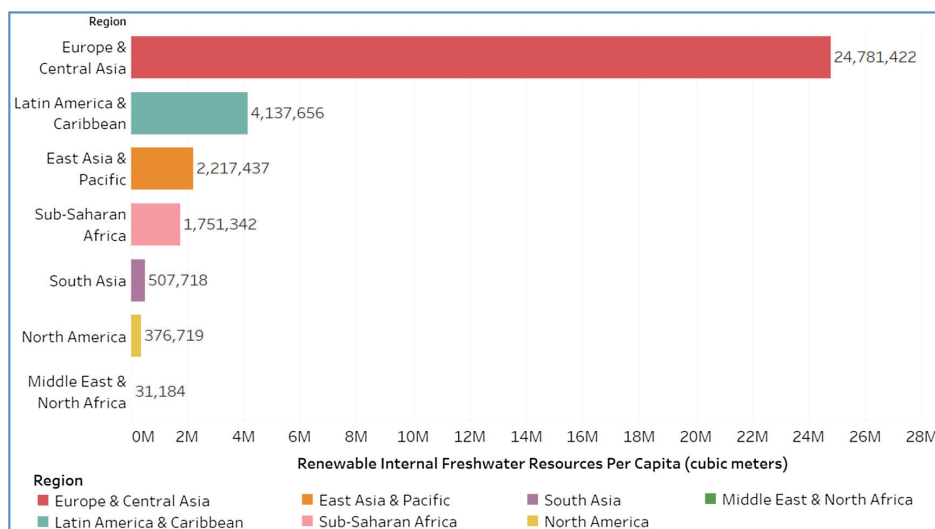
**Fig. 8.** United States' Total Greenhouse Gas Emissions

The second variable that was examined in this study was Population by Region. From the following bar chart, it is shown that East Asia and the Pacific has the highest population compared to any other region, followed by South Asia, and then by Sub-Saharan Africa. The region with the lowest population compared to any other region is North America, followed by the Middle East and North Africa.



**Fig. 9.** Total Population by Region

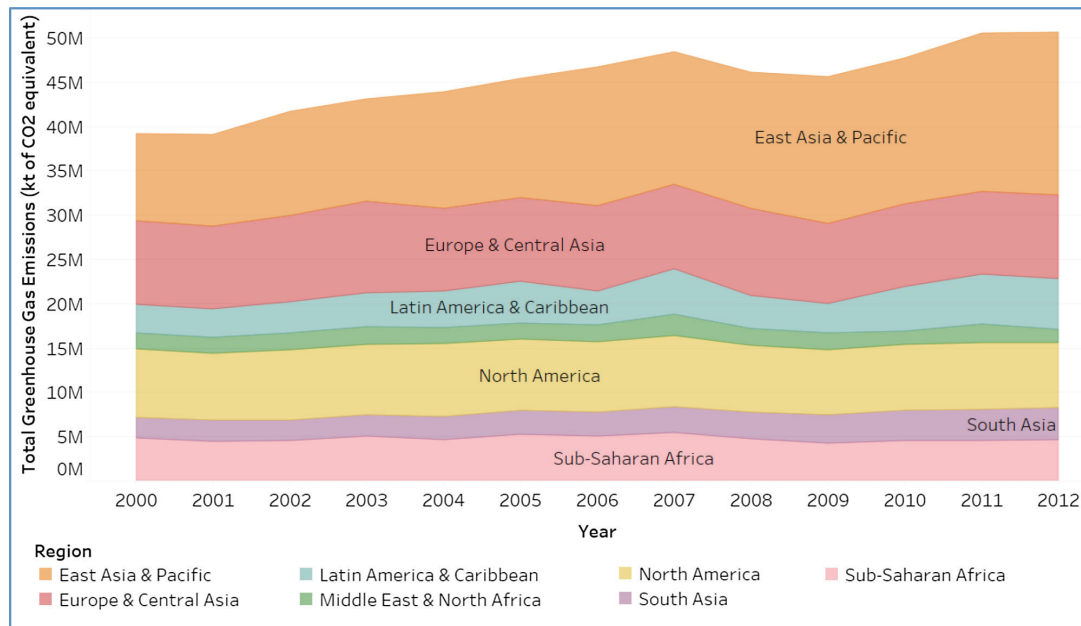
The next environmental factor examined in this study was the Renewable Internal Freshwater Resources per Capita (in units of cubic meters). From the following bar chart, it is shown that the region with the highest amount of renewable internal freshwater resources per capita is Europe and Central Asia. The region with the lowest amount of renewable internal freshwater resources per capita is the Middle East and North Africa, followed by North America, and then by South Asia. Sub-Saharan Africa also has a low amount of renewable internal freshwater resources per capita.



**Fig. 10.** Renewable Internal Freshwater Resources per Capita by Region



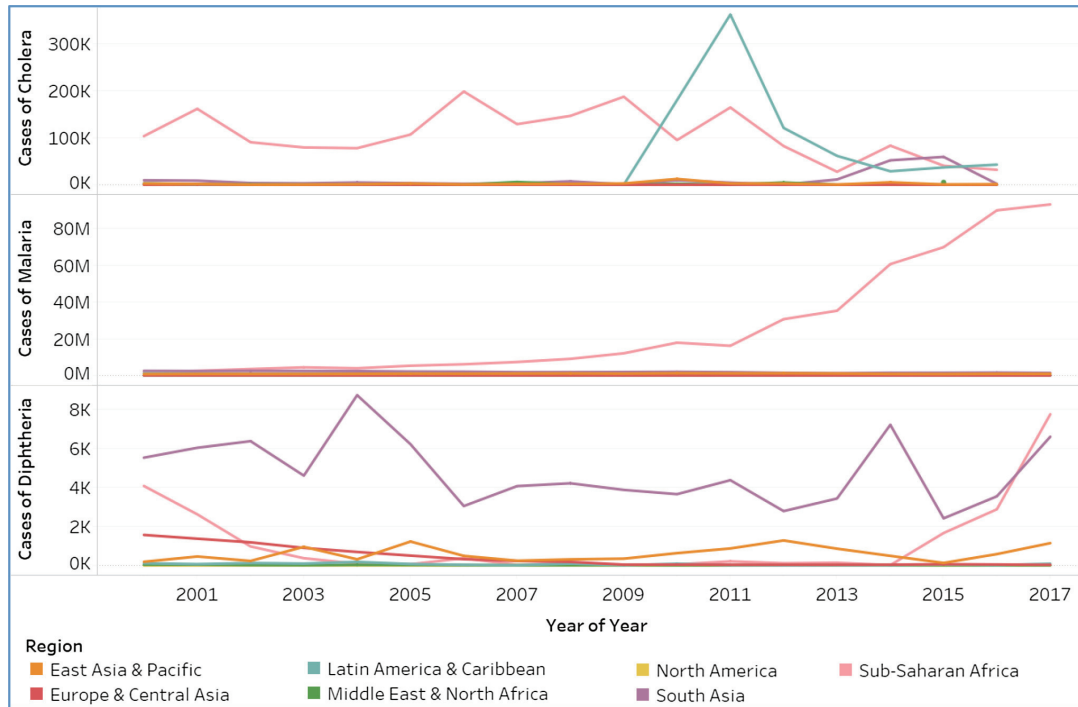
The final environmental that was examined was Total Greenhouse Gas Emissions over Time by Region. Figure 11 shows an area chart that was created in Tableau, and it can be seen that East Asia and the Pacific has the highest amount of total greenhouse gas emissions, followed by Europe and Central Asia.



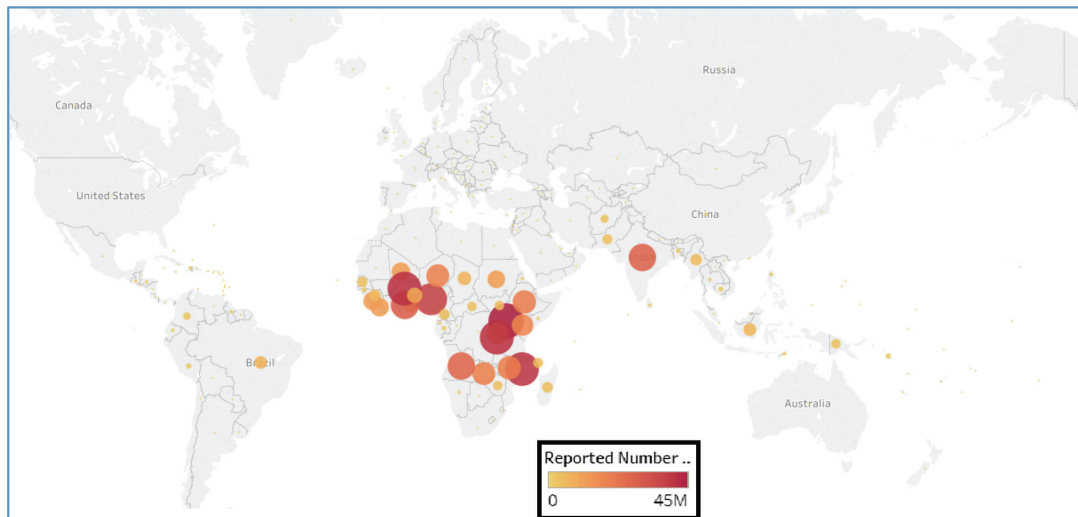
**Fig. 11.** Total Greenhouse Gas Emissions over Time by Region

After examining some of the environmental data, the health data gathered from the Global Health Observatory [10] was then examined. Figure 12 shows a line chart of the Number of Reported Cases of Communicable Diseases over Time by Region. The communicable diseases focused on in this study were Malaria, Cholera, and Diphtheria. From the line chart, it is shown that Malaria has the highest number of reported cases, in the millions, compared to the number of reported cases of Cholera and Diphtheria, which are only in the hundreds of thousands and thousands, respectively.

Because malaria had the highest number of reported cases compared to cholera and diphtheria, this disease was further looked into. In Fig. 13, another symbol map was created to show which regions or countries had the highest number of reported cases of malaria. It is shown that Africa is the region with the highest number of reported cases of malaria, as well as the country of India.

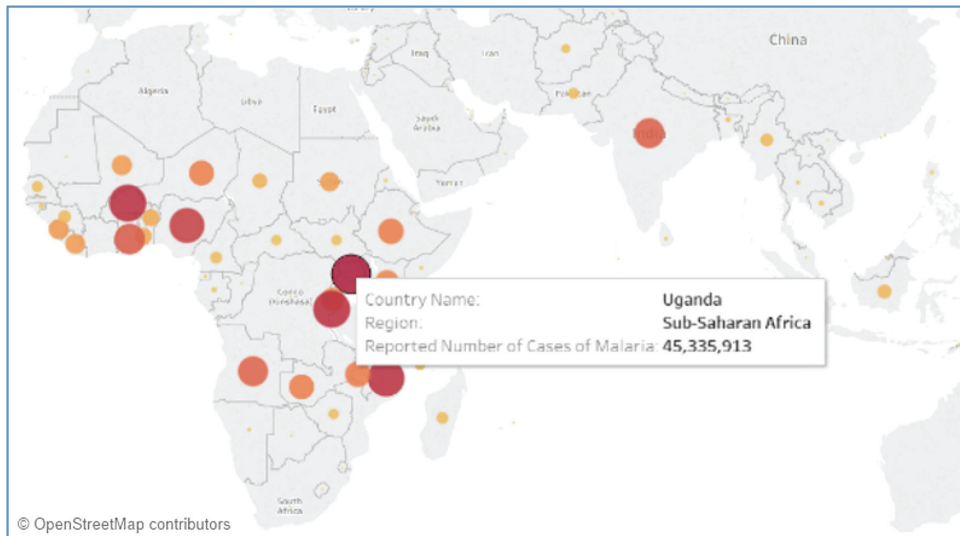


**Fig. 12.** Number of Reported Cases of Communicable Diseases over Time by Region

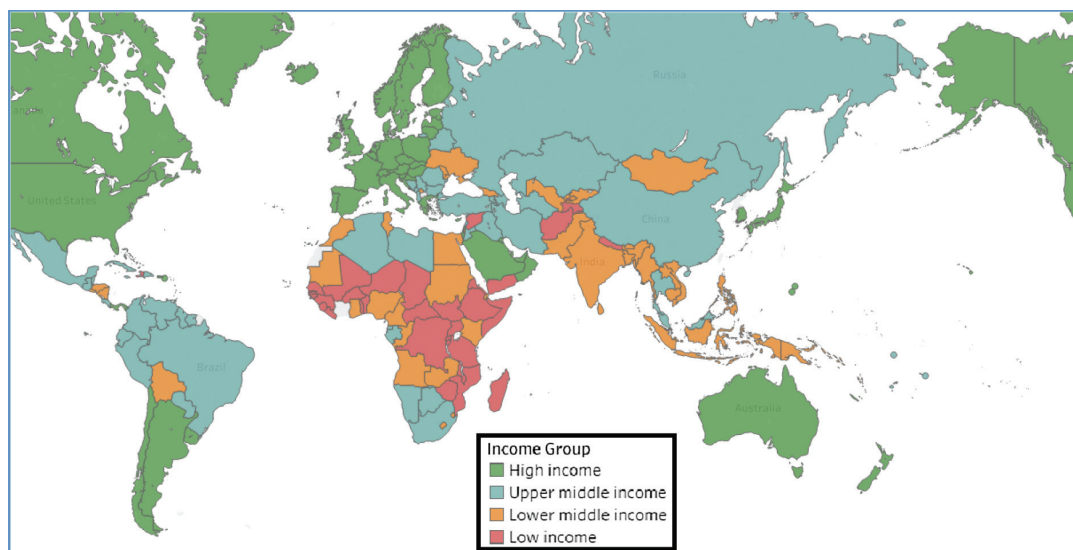


**Fig. 13.** Number of Reported Cases of Malaria by Country

The final factor that was examined in this study was the Income Group of each Country. Figure 15 is a map that was created in Tableau which shows which income group each country belongs to. The income groups were categorized as high income, upper middle income, lower middle income, and low income. Much of Africa falls into the low income to lower middle income group, with a few upper middle income



**Fig. 14.** Number of Reported Cases of Malaria in Africa and South Asia



**Fig. 15.** Income Group for each Country

countries in the north and south of Africa. The regions with the highest income groups are Australia, Europe, and most of North America with the exception of Mexico, which falls under the upper middle income group.

## 4 Conclusions

The charts and maps that were created using Tableau show that the Number of Reported Cases of Communicable Diseases, specifically Malaria, may be related to both the Number of Internal Freshwater Resources per Capita, and the Income Group of each country. When comparing Fig. 10 (Renewable Internal Freshwater Resources

per Capita by Country) to Fig. 13 (Number of Reported Cases of Malaria by Country), it can be concluded that the amount of renewable internal freshwater resources a country has may play a role in the number of reported cases of malaria for that country. Sub-Saharan Africa is one of the regions with the lowest amounts of renewable internal freshwater resources per capita, and is also has the highest number of reported cases of malaria. Additionally, South Asia, the region with the third lowest amount of renewable internal freshwater resources per capita, also has a high number of reported cases of malaria, especially in India.

In Fig. 10, it is shown that North America is the region with the second lowest amounts of renewable internal freshwater resources per capita. However, this region has very little to no cases of malaria, even though most of the other regions, such as Sub-Saharan Africa and South Asia, which also have low amounts of renewable internal freshwater resources per capita, have high cases of malaria. The difference between these regions may be due to the income level of the countries in those regions.

In Fig. 15 (Income Group for each Country), it is shown that most of the countries in Africa fall under the Low Income Group and Lower Middle Income Group. These countries are also the countries, as seen in Fig. 14, which have high numbers of reported cases of malaria. The countries in Africa which fall under the Upper Middle Income Group do not have the same high number of cases of malaria as the countries in the lower income groups. The cases of malaria in these upper income level countries are relatively low compared to the lower income level countries in the same continent. Additionally, North America, which has low amounts of renewable internal freshwater resources per capita, falls under the High Income Group. It appears that the high income level of this region may contribute to it not having a high number of cases of malaria as Africa does. The high income level of this region may allow the people who live in this area to be better equipped at preventing the spread of these communicable diseases, as well as treating any cases of these diseases if they should occur.

A final conclusion that can be made from this study is that air pollutions, specifically greenhouse gas emissions, do not seem to play a role in the prevalence of the three communicable diseases examined in this study. In Figs. 1 and 5, it is shown that East Asia and the Pacific, North America, and Europe and Central Asia all have high amounts of total greenhouse gas emissions, yet these regions have very little to no cases of malaria, cholera, or diphtheria. If this study were to be continued, another health factor that can be examined is non-communicable diseases, in addition to the communicable diseases that have already been examined.

Although this study showed that the prevalence of malaria, cholera, and diphtheria are highest in the regions where renewable internal freshwater resources are low, and in the countries where the income group is low, this does not necessarily imply causality between these factors. More research would have to be done in order to determine the causality between the environmental factors and the health factors used in this study.

## References

1. Landrigan, P.J., et al.: The lancet commission on pollution and health. *Lancet Comm.* **391**, 462–512 (2017). [https://doi.org/10.1016/S0140-6736\(17\)32345-0](https://doi.org/10.1016/S0140-6736(17)32345-0)
2. Ban, J., et al.: Environmental health indicators for china: data resources for Chinese environmental public health tracking. *Environ. Health Persp.* **127**(4), 1–10 (2019). <https://doi.org/10.1289/EHP4319>
3. Sun, Z., Zhu, D.: Exposure to outdoor air pollution and its human health outcomes: a scoping review. *PLoS ONE* **14**(5), 1–18 (2019). <https://doi.org/10.1371/journal.pone.0216550>
4. Setty, K., et al.: Faster and safer: research priorities in water and health. *Int. J. Hyg. Environ. Health* **222**(4), 593–606 (2019). <https://doi.org/10.1016/j.ijheh.2019.03.003>
5. Pickering, A.J., Davis, J.: Freshwater availability and water fetching distance affect child health in Sub-Saharan Africa. *Environ. Sci. Technol.* **46**(4), 2391–2397 (2012). <https://doi.org/10.1021/es203177v>
6. World Health Organization: Health & environment: tools for effective decision-making (2005). <https://www.who.int/heli/publications/brochure/en/>
7. Wei, L., Feeny, D.: The dynamics of the gradient between child's health and family income: evidence from Canada. *Soc. Sci. Med.* **226**, 182–189 (2019). <https://doi.org/10.1016/j.socscimed.2019.02.033>
8. Torras, M.: The impact of power equality, income, and the environment on human health: some inter-country comparisons. *Int. Rev. Appl. Econ.* **20**(1), 1–20 (2006). <https://doi.org/10.1080/02692170500362199>
9. World Health Organization: Global Health Observatory (GHO) data (2019). <https://www.who.int/gho/en/>
10. The World Bank: World Bank Open Data (2019). <https://data.worldbank.org/>