Heart Disease around the World: Exploration and Analysis of Factors and Trends

Changzhou Li
Xueyuan Li
Minwoo Sohn
Amogh Vig

Contents

Introduction

The overarching aim of our project is to investigate heart disease in different countries from different key standpoints. Specifically, we will analyze this issue in-depth by examining the local climate, economy, and political system. The question we hope to answer is if these factors are related to the mortality rate for heart disease. This will involve combining multiple different data sets to make a single data frame on which we can perform our analysis and inference. To accomplish this, our project is divided into three main components, or phases, as presented in the following sections. We chose to focus on heart disease since preliminary research showed that this is the #1 killer worldwide. We corroborate this information using our own data and analysis in Phase 1. Also, please note that the data for Phases 1 and 2 is only for the year 2019, since this was the most recently available data that we could find for most of the relevant variables. However, the data in Phase 3 spans several decades because we conclude our study by conducting a time series analysis for heart disease death rate in countries of interest to identify current trends and make predictions about future heart disease mortality rates.

Data Collection Process

The data used in this project was stored as tables in a variety of different online sources. In order to get the data into a form that we could manipulate, we either downloaded the data directly as a CSV file from the website or scraped the data into an R data frame using the `read_html` function from the `rvest` package. The raw data required some cleaning to be used for our analyses. This cleaning was accomplished through conditional R programming, which was used to remove rows with missing values, remove extraneous unwanted variables, and select the relevant observations. The cleaned data was then used to generate all the figures and analyses presented in this report. Furthermore, in our study, we use linear regression multiple times to study the relationship between two quantitative variables. The correlations we present for these analyses were calculated using R's `cor` function. Therefore, a combination of web scraping and data manipulation using helpful functions available from R libraries allowed us to extract and transform our data into the required format to create this report.

Variable Transformations

For some of the analyses in this report, we had to generate new variables by transforming the existing variables in the raw data to discover new insights. These variable transformations are described and presented below:

1. The goal of this project is to study factors and trends that are related to heart disease on a global scale. For this analysis, it is much more appropriate to consider the mortality rates due to heart disease in different countries or regions, rather than considering the raw death counts. This is because the rate of heart disease mortality accounts for differences in population, which allows for a fair comparison between different countries or regions. In light of this, we generated a new variable that captures the heart disease death rate for different countries. This was accomplished through a simple transformation of two of the raw data variables: we divided the total number of deaths due to heart disease by the population of the country, thereby computing the proportion of the population who died due to heart disease that year.

2. One of our objectives was to find connections between heart disease mortality and economic indicators. By far the most widely used measure for comparing different economies is gross domestic product (GDP). However, similar to the raw number of deaths due to disease, GDP does not account for differences in population and thus is not conducive to fair comparisons of economic prosperity. A better measure to use for this purpose is GDP per capita, which conveys how much wealth the average citizen of that country has, rather than simply being the gross sum of the entire wealth of the country. Once again, this new variable was generated through a simple transformation of the raw data variables:

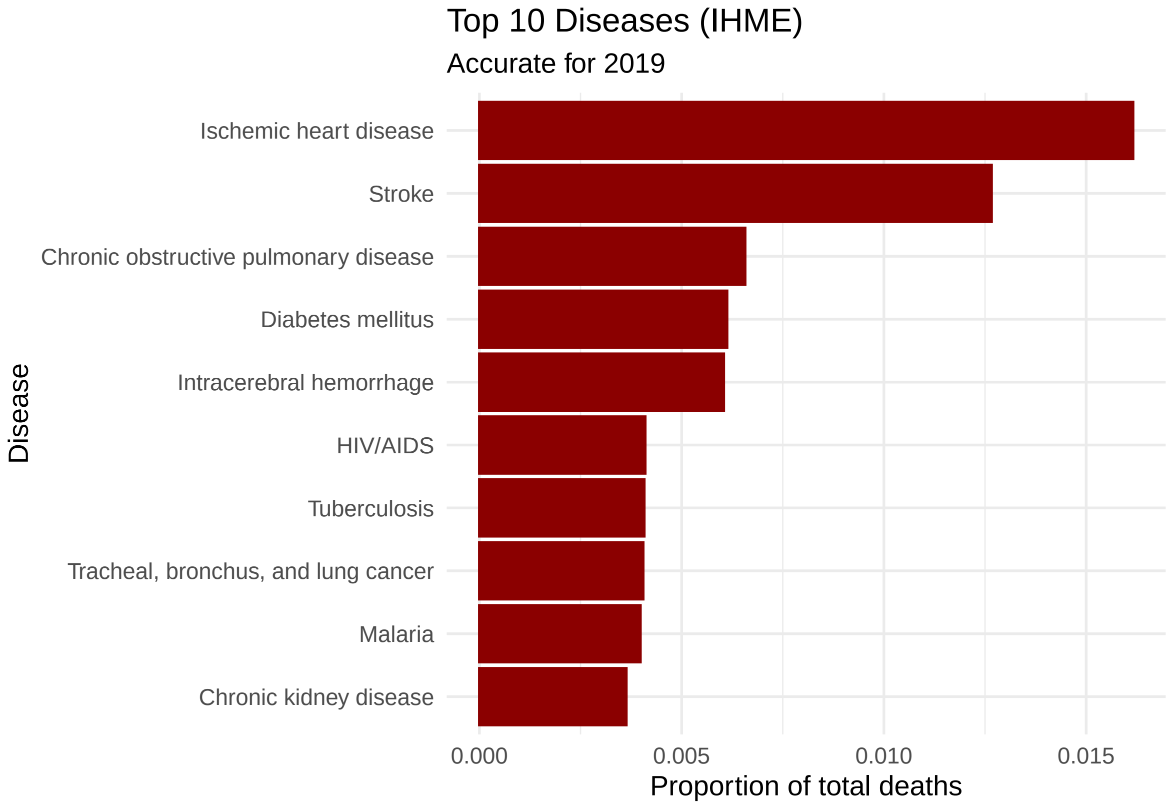
we divided the total GDP of the country by the population of the country, thereby computing the GDP per person, which is the same as the GDP per capita.

3. When looking at the relationship between heart disease mortality and the climate, we first wanted to see if hot and cold countries show a significant difference in mortality rates. To do this, we first had to decide how to classify countries as being hot or cold. The mean annual average temperature in our data set is approximately 20.9°C. Therefore, for this study we define countries that have an average annual temperature of 20.9°C or lower as being cold countries, while countries that have a higher average annual temperature are considered to be hot countries. Once this was established, the new variable used to classify countries as hot or cold (temperature category) was generated by calling the `mutate` and `ifelse` functions in R on the raw climate and temperature data.

Phase 1: Disease and mortality

The first phase of our project will be to analyze disease and mortality data. The goal is to identify the current deadliest disease (the disease that has been responsible for the most deaths worldwide, based on the most recent data available). This disease will be our focus for the remainder of our analysis. Preliminary research showed that heart disease kills the most people globally. We wish to verify this information using our own analysis.

The data used for this phase comes from the official website of the Institute for Health Metrics and Evaluation (IHME) [12]. This data contains the total global mortality rates for different diseases for the year 2019. However, before this data could be used, it first had to be cleaned. This is because our analysis requires looking at the mortality rate for only individual diseases, but the raw data included some observations (rows) that aggregated mortality rates for a family of diseases. Examples include respiratory infections and neoplasms. Additionally, the raw data included some causes of death that are not actually diseases, such as interpersonal violence, defects during pregnancy and birth, and trauma. Incorporating these rows into the analysis would lead to misleading results, since we wish to represent the deadliness of individual diseases, and not groups of diseases or other causes of death. Therefore, conditional R programming was utilized to subset the raw data such that these unwanted rows were removed. The resulting data was used to generate a bar plot to show different diseases and the proportion of total deaths that they were responsible for in the year 2019. The resulting plot is shown below.



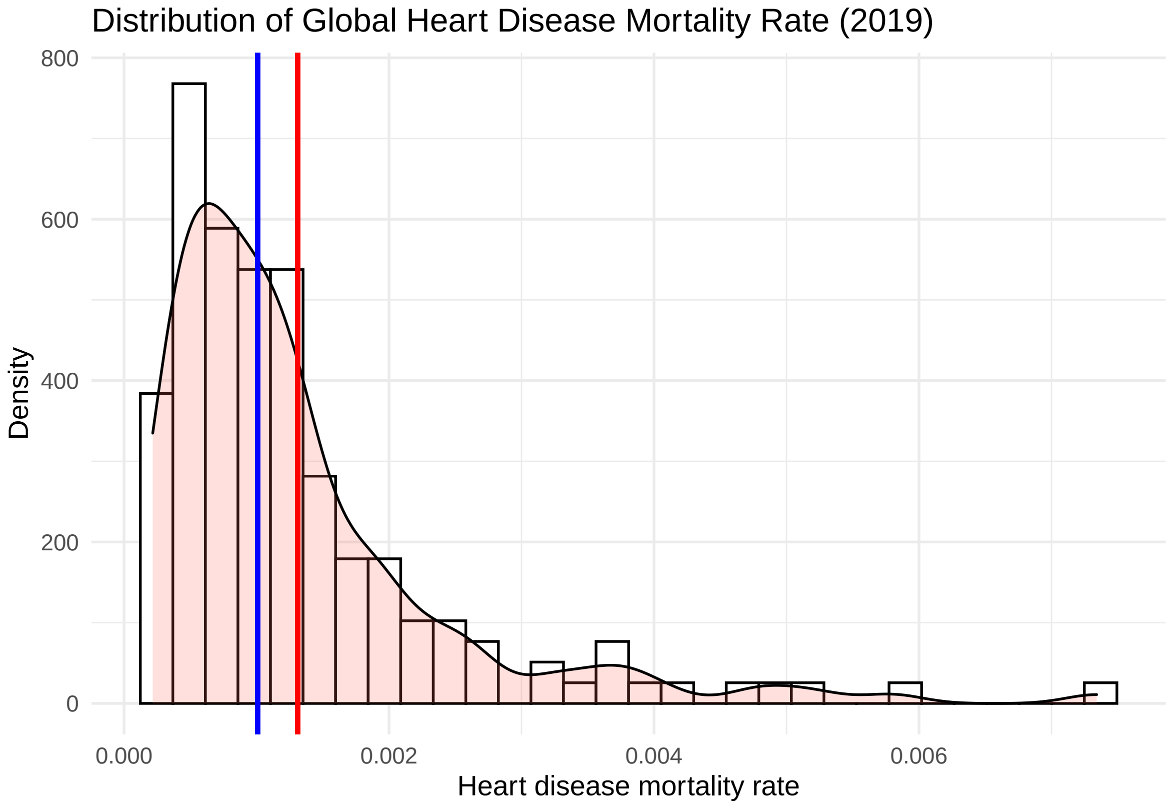
The resulting plot shows that ischemic heart disease, also knows as coronary heart disease (CHD), was the biggest killer worldwide for the year 2019, accounting for around 1.6% of total deaths from disease worldwide. This highlights the need for a deeper collective understanding about the factors that contribute towards death from heart disease. Work in this field could help to significantly reduce the loss of human lives on a global scale. Thus, Phases 2 and 3 of this project will focus on ischemic/coronary heart disease.

Phase 2: Main Exploratory Data Analysis

Phase 2 represents the main body of our project. We will explore the relationship between heart disease mortality rate and several environmental, economic, and political factors. We wish to determine if there are any interesting trends or correlations that may not be immediately obvious.

Distribution of Heart Disease Mortality Rate

Before diving into the analysis, it is useful to first examine the global distribution of heart disease mortality rates. This may be accomplished using an overlaid histogram and density plot.

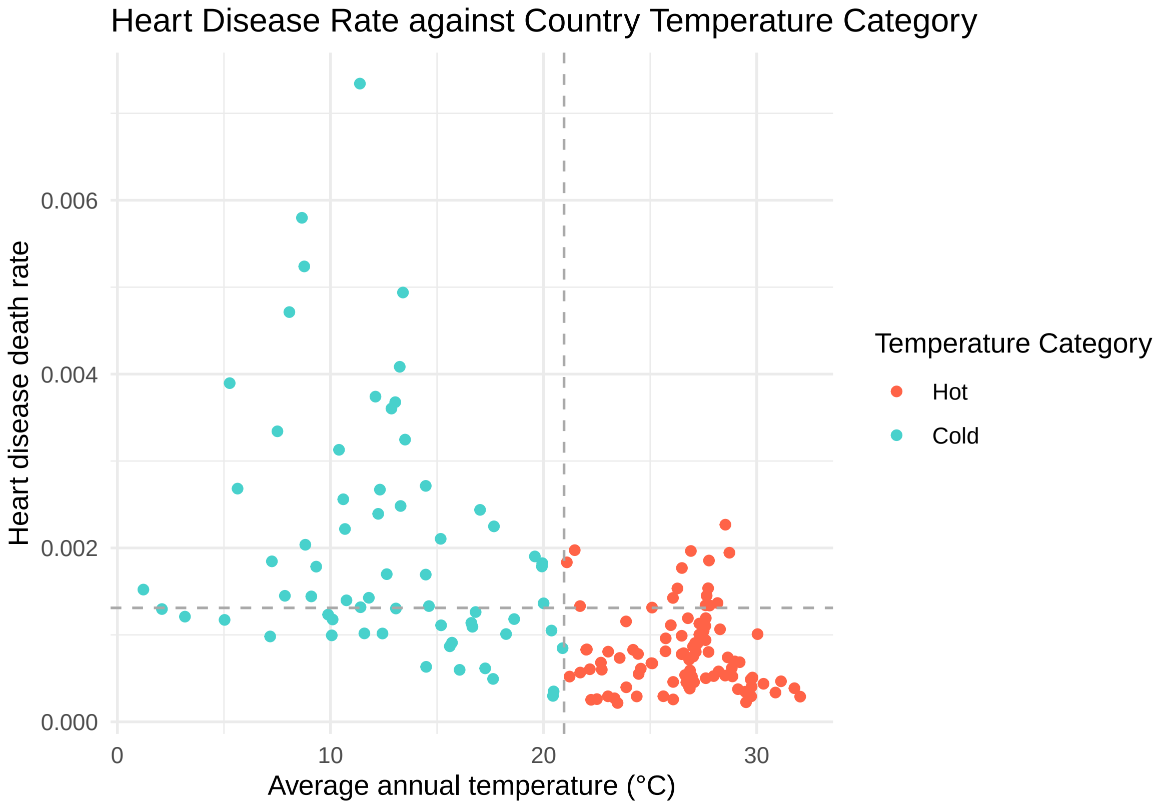


From this plot, it is evident that the distribution of global heart disease mortality rate is heavily skewed right (positive skewness). This means that the mean heart disease mortality rate is higher than the median heart disease mortality rate, which in turn is higher than the mode of the heart disease mortality rate. This is further clarified by the inclusion of the vertical lines, where the mean is shown in red and the median is shown in blue. The mode of the heart disease mortality rate is represented by the peak of the histogram and density plot. The key takeaway is that lower rates of death from heart disease are more common globally than higher rates of death from heart disease.

Heart Disease and Climate

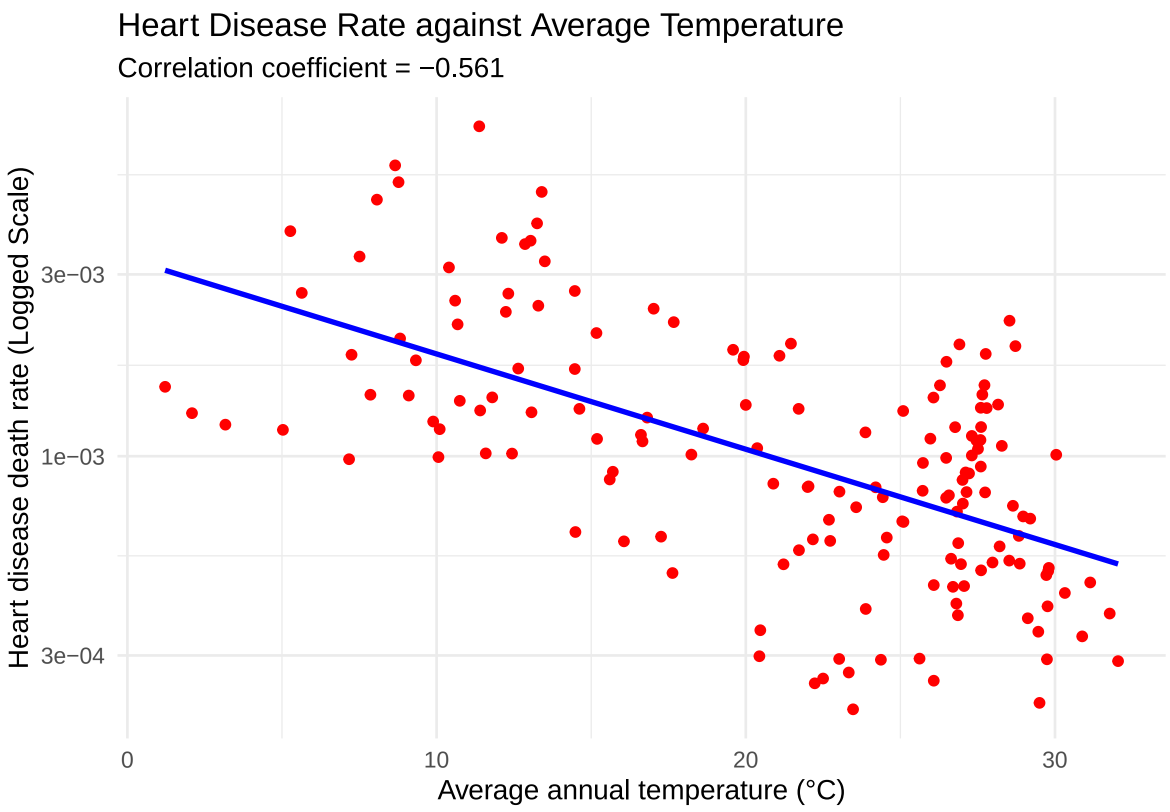
Heart Disease and Average Annual Temperature

Our goal is to establish if there is a relationship between deaths from heart disease and average annual temperature. As a first step, we will divide countries into two categories: hot climates and cold climates. By visualizing the binary outcome, we may gain some intuition about how average annual temperature and heart disease mortality rates may be linked. Dividing up the country temperature in this manner yields the following plot.



In this visualization the mean average annual temperature and mean heart disease death rate have been represented by the vertical and horizontal dashed lines respectively. Thus, by looking at the number of points above the horizontal dashed line, it is made evident that cold countries appear to suffer from much higher rates of heart disease than hot countries. This provides some indication that heart disease deaths and average annual temperature may be connected in some way.

Let us explore this issue further by evaluating the relationship between heart disease death rates and average annual temperature numerically. This can be accomplished by visualizing the relationship using a scatter plot and performing linear regression on the data to obtain a line of best-fit, which will allow us to measure the strength of the relationship between between these two variables by calculating the correlation between them.



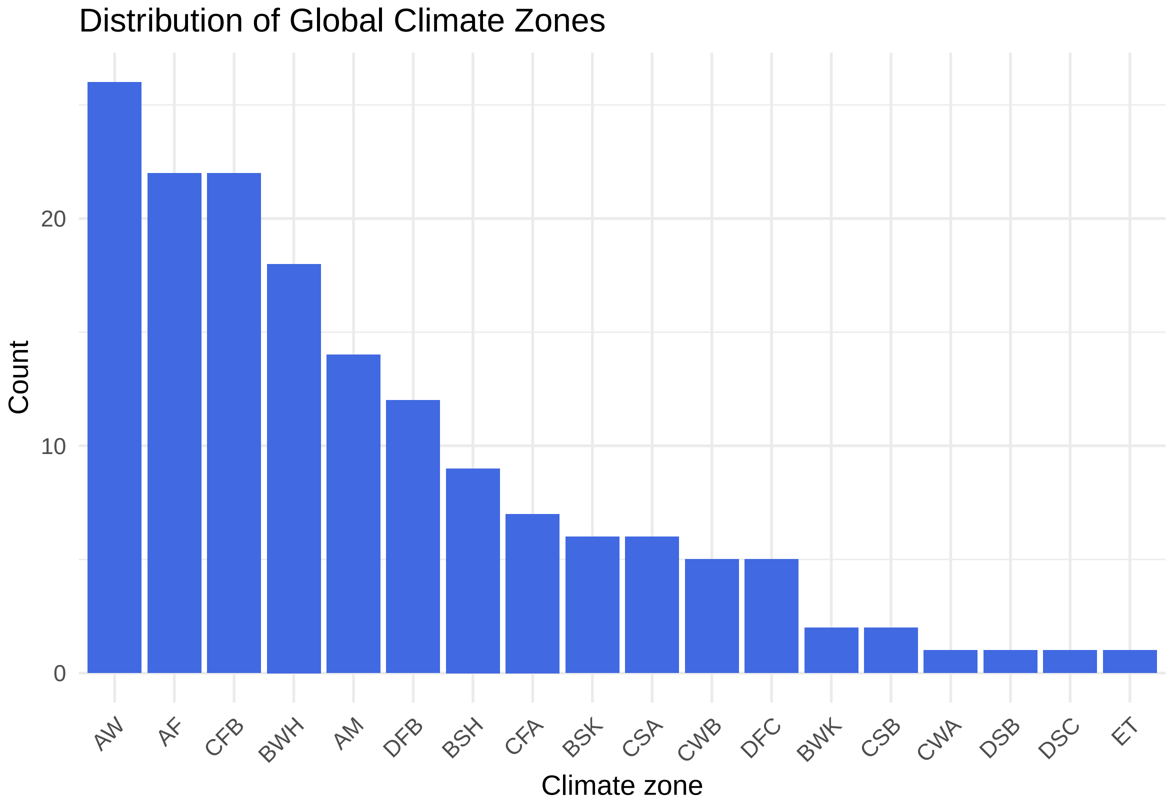
The resulting plot appears to show a weak negative correlation between deaths from heart disease and average annual temperature. The R^2 value for this correlation is around -0.561. This means that only about 56.1% of the variation in heart disease death rate for our data can be explained using this linear model.

These results mean that the relationship between these two variables is sufficiently weak that it is doubtful whether this relationship truly exists or is simply present in our data by coincidence. After all, this data is representative of only 1 year, 2019, so it is possible that this observed weak correlation is a fluke, and that in reality no such correlation exists. Nonetheless, we can conclude to a high degree of confidence that, for the year 2019, the average annual temperature was not a reliable predictor of heart disease death rate. However, we still have evidence that suggests that colder countries have higher rates of mortality from heart disease than warmer countries. Therefore, we must consider other variables. If temperature is not strongly correlated with heart disease mortality rate, then perhaps other environmental factors that are shared by colder countries, such as the specific type of climate, may better explain our findings.

Heart Disease and Climate Type

The next step in this phase of the analysis is to study the relationship between deaths from heart disease and different climate zones. For this project, we use the Köppen climate classification of climate zones [5]. For brevity and convenience, only the abbreviations of these different climate categories are stored in our data, and thus only these codenames appear in our figures. Please refer to the provided data dictionary for the full names of these different climate types.

To begin, it would be appropriate to first visualize the distribution of the different climate types that are present in our data to familiarize ourselves with this variable. This can be done using a bar plot, as shown below.

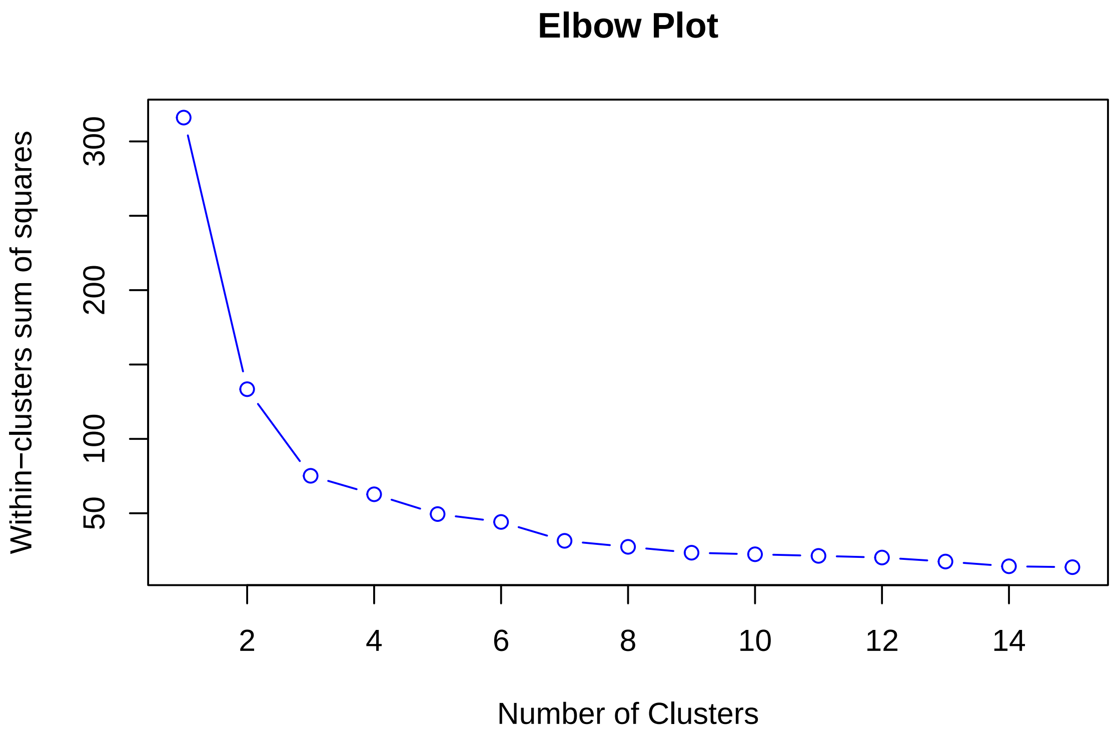


The resulting plot shows that the most common climate type in our data set is AW, which is tropical savannah. The rarest climate types in our data set are CWA (dry-winter humid subtropical), DSB (warm, dry-summer continental), DSC (dry-summer subarctic), and ET (tundra). There is only one instance for each of these climate types in our data.

Visualizing the distribution of climate zones does not reveal any information about how climate zones may be related to heart disease mortality rate. A different visualization is needed for this. Since climate zone is a categorical variable, any relationship between climate zone and heart disease mortality rate that could exist cannot be discerned using linear regression, since linear regression is only applicable for numerical (quantitative) variables. Instead, we will need to use a different algorithm to obtain and visualize these results. Therefore, we use a K-Means clustering algorithm to group our data into distinct clusters. This will allow us to make a new visualization to determine if our heart disease death rate data can be divided into different groups of climate types.

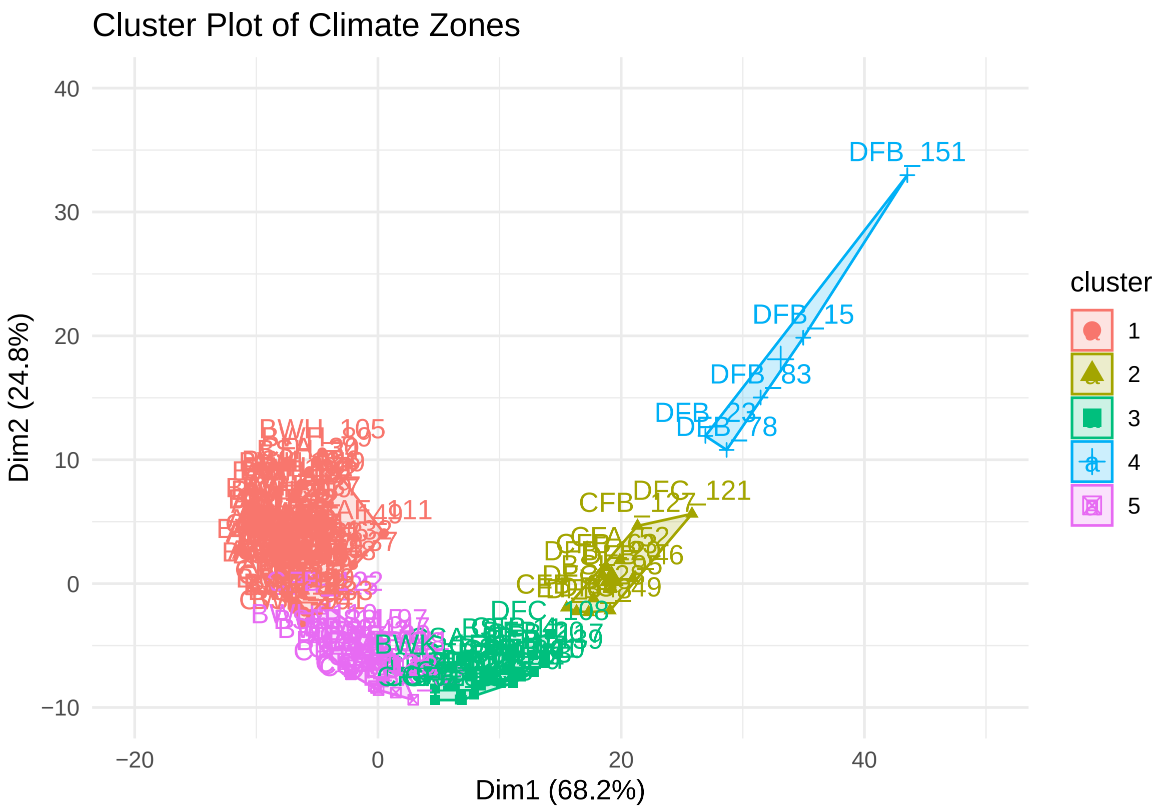
For our K-Means clustering algorithm, the three variables we use are the climate zone, the average annual temperature, and the heart disease mortality rate. The climate zone variable acts as the ground-truth labels for the data, while the average temperature and heart disease mortality rate values are fed into the K-Means algorithm as unlabeled data. Once the clustering is complete, the model performance can be evaluated by determining the total variance in the data set that is explained by the clustering (similar to how the R^2 value represents the total variance explained by linear regression).

Before conducting K-Means clustering on our data, we must first determine the optimal number of clusters to use for the algorithm. This can be accomplished with the aid of an elbow plot. The elbow plot graphs the within-clusters sum of squares against the number of clusters for the given data. Within-clusters sum of squares is a monotonously decreasing function. Thus, by identifying the threshold where decreasing the within-clusters sum of squares further does not yield significant improvements in the clustering performance (i.e. the “elbow” of the plot), we can find the ideal number of clusters to use. The elbow plot for our data is presented below.



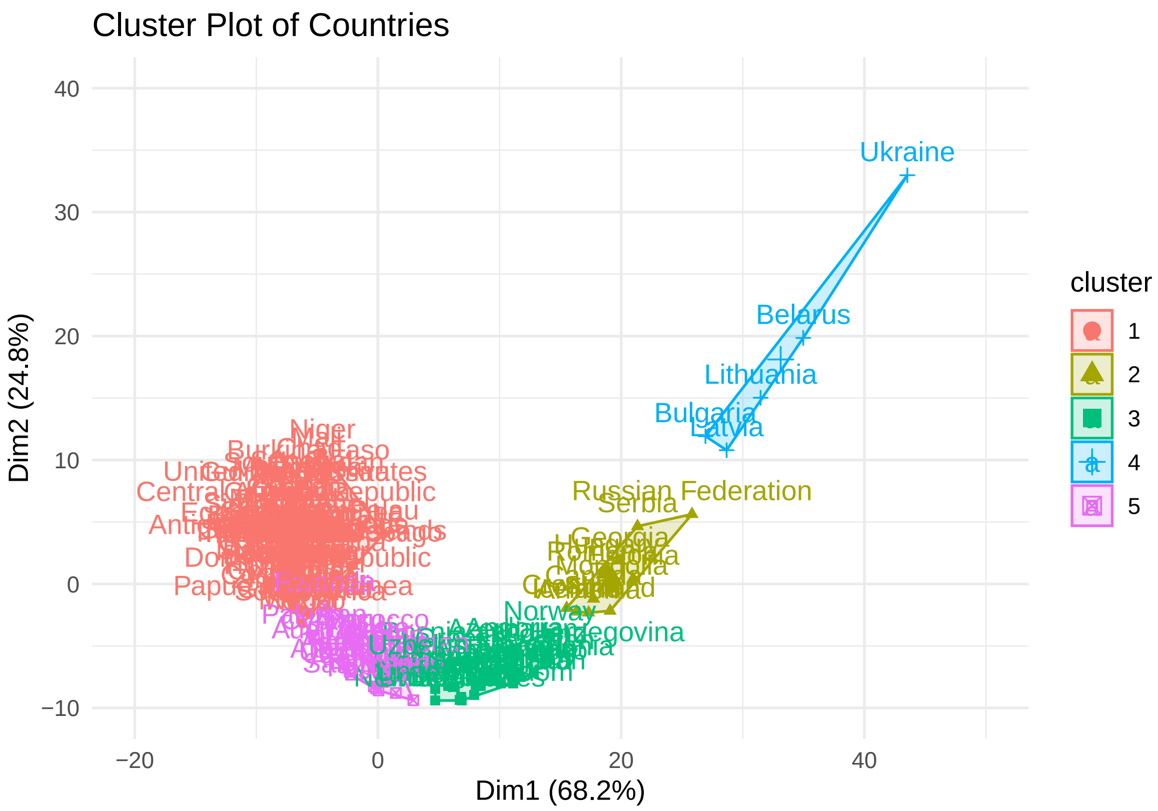
From this elbow plot, it is evident that increasing the number of clusters beyond 5 doesn't lead to significant improvements. The within-clusters sum of squares begins to plateau after this point. Therefore, we will use 5 clusters for our K-Means clustering algorithm.

Performing K-Means clustering on the average temperature and heart disease mortality rate data yields the following plot.



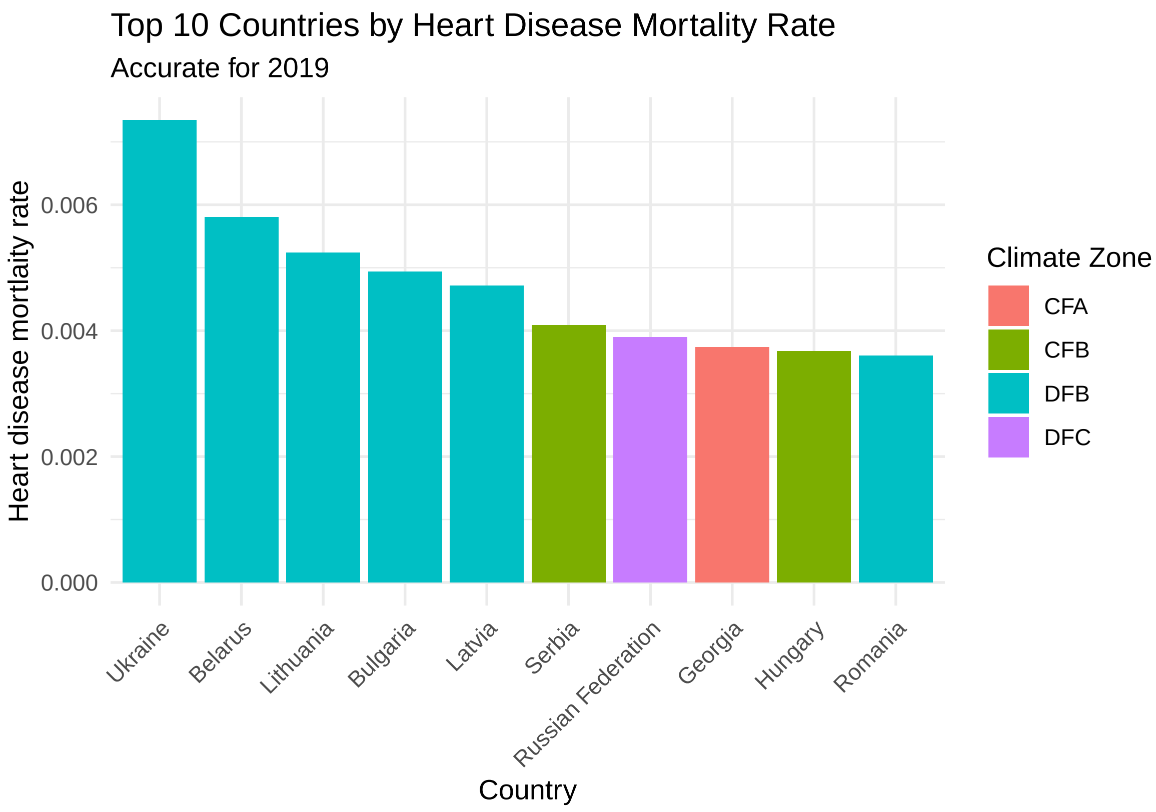
In the resulting plot, the data has indeed been grouped into 5 different clusters. Together, the clusters form a characteristic hockey-stick shape. Most of the clusters contain different types of climate zone within them. However, cluster number 4, shown in blue, is an exception and is therefore of particular interest. Notice that this cluster only contains points that represent the DFB climate zone, which corresponds to cold continental climates without a dry season. Note also that our K-Means algorithm was able to explain 87.7% of the total variance in the data set by classifying the data into these clusters. This indicates a good fit, allowing us to be confident in the clusters that K-Means has found.

Given that the data points in cluster 4 all correspond to countries that have the same climate, it is reasonable to assume that they are geographically close to each other. We can verify is by repeating the K-Means clustering, but using countries as labels rather than climate zones. This yields the plot below.



From this visualization, we can identify the countries in cluster 4. These are all cold countries as per the given definition for this analysis. Furthermore, it appears that all 5 countries in cluster 4 are eastern European countries. In fact, these are the countries that have the top 5 highest heart disease mortality rates in our data.

Our findings in this section are summarized in the graph below, where we display the countries with the 10 highest rates of heart disease mortality for the year 2019.

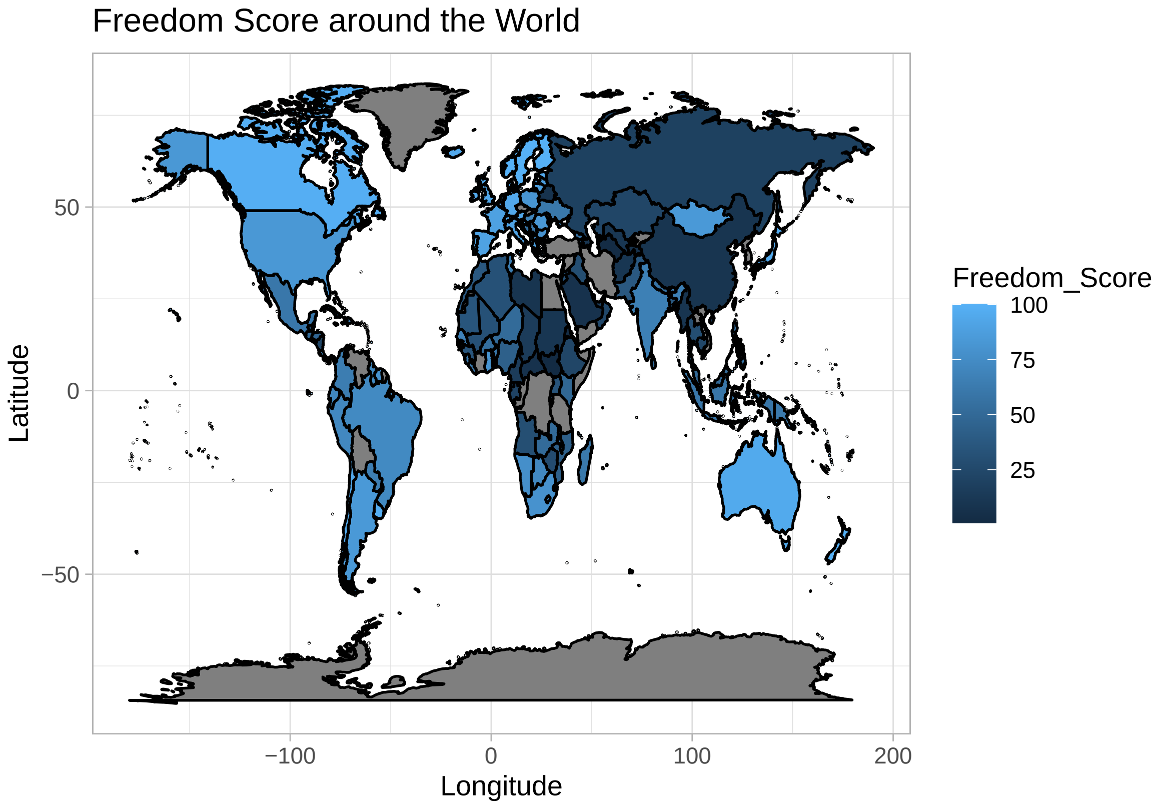


This bar plot shows that of the countries with highest heart disease mortality rates, 90% are located in eastern Europe (the only exception is Georgia) and 60% have a cold continental climate without a dry season (DFB). Note also that the five countries with the highest heart disease death rates all have a DFB climate. Given the geographic proximity of the majority of these countries, it is reasonable to suggest that they share factors in common that may be good predictors for heart disease mortality rate, such as culture (for example, an active lifestyle versus a more sedentary lifestyle) or diet. Perhaps political and economic factors may also be at play. These are analyzed in the following sections.

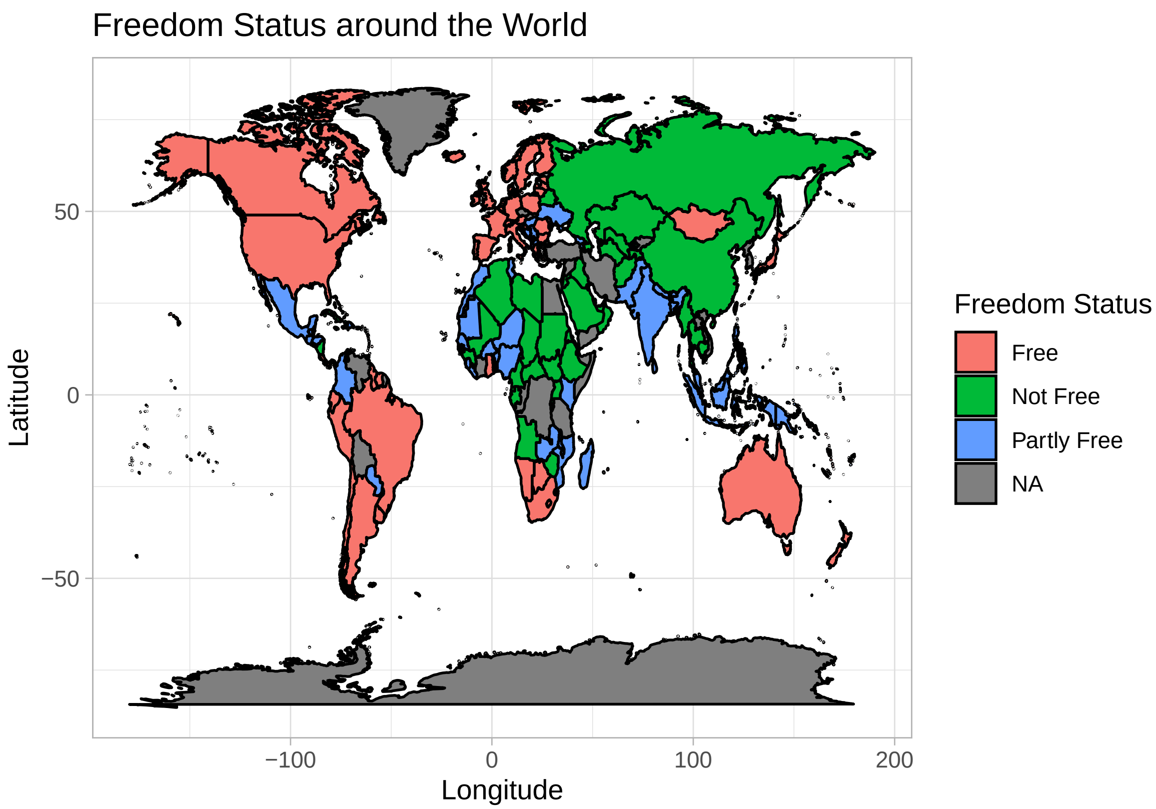
Heart Disease and Freedom

Freedom Score and Status around the World

The political variables we focus on for this project are concerned with the different levels of freedom in different countries. Before finding the relationship, we first need to understand the metrics of freedom. The Freedom score measures access to political rights and civil liberties among 210 countries and territories on a scale of 0 to 100. To be more specific, the political rights are evaluated based on the electoral process, political pluralism and participation, and government functioning under 10 sub-questions that are worth 4 points each, adding up to a total of 40 points [7]. Civil liberties are determined by measuring freedom of expression and belief, organizational rights, rule of law, personal autonomy and individual rights. There are 15 sub-questions related to the topics mentioned above that are worth 4 points each, adding up to a total of 60 points [7]. Thus, the two categories sum up to a maximum of 100 points. Freedom Status divides the numerical Freedom score into 3 categories, “Free”, “Partly Free”, and “Not Free”, thereby providing a meaningful interpretation of the score.



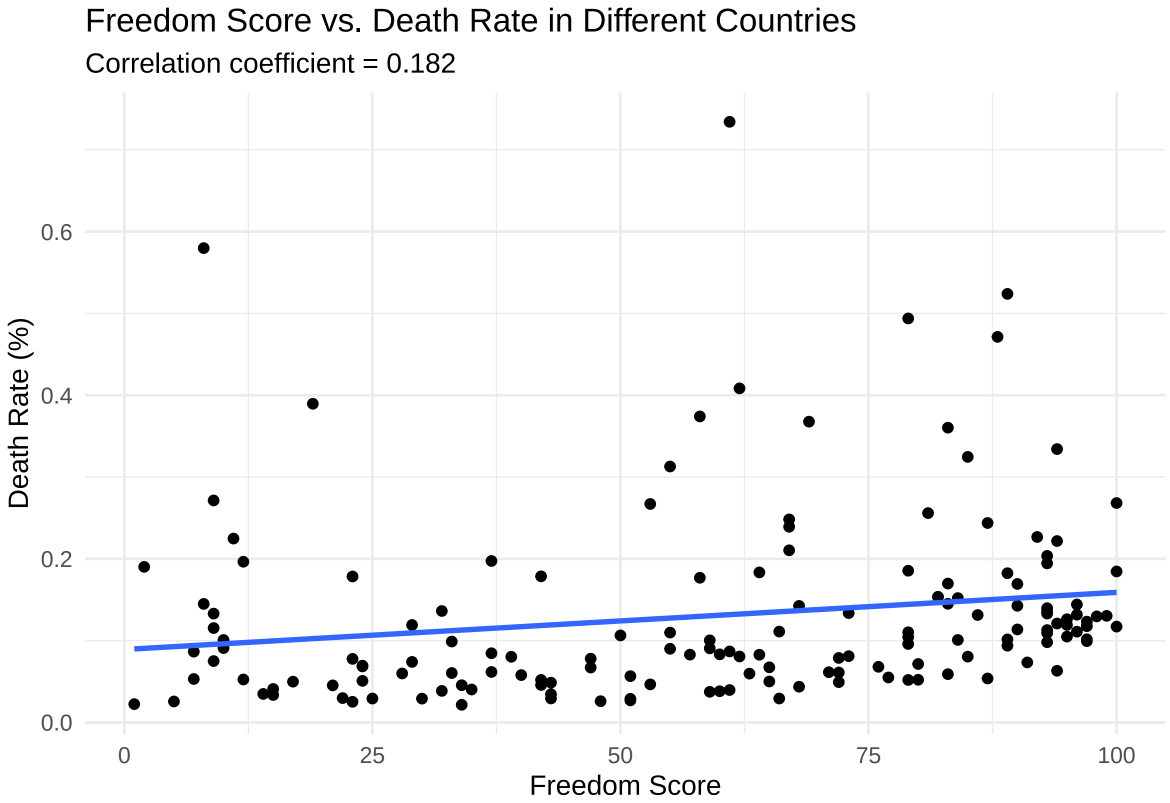
The first world map is colored based on the Freedom Score. Darker colors represent the countries with low freedom scores, and lighter colors depict the countries with higher freedom scores. The countries with higher freedom scores are mostly located in North and South America, Europe, Oceania, and East Asia. In contrast, the countries with low freedom scores are in central Africa, central Asia, and Southeast Asia.



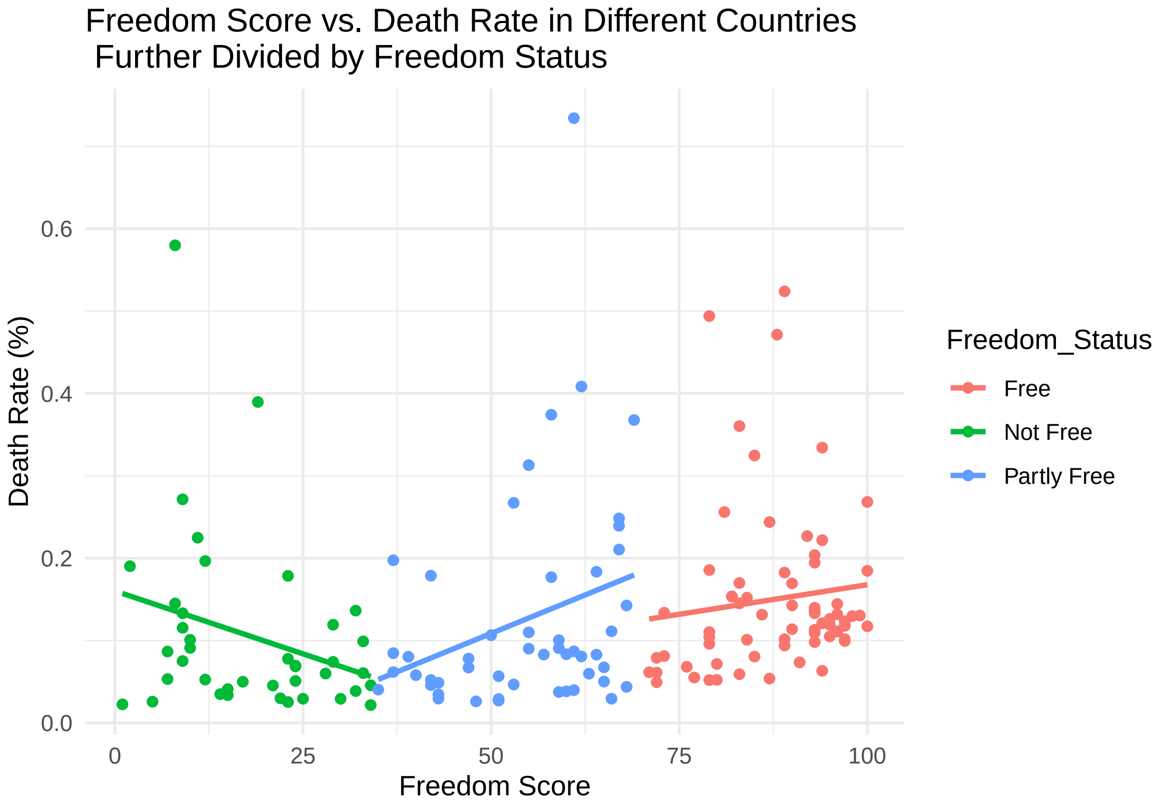
The second world map is colored on 3 freedom statuses to give better distinguish the countries. Summing up the result, we can observe that most of the western world, countries in North America, South America, and Western Europe have a freedom score above 70 and are considered free. Countries in Southern Africa, Oceania, and East Asia also fall within the free category. Central American countries, East and Western African countries, Southern Asia, and Southeast Asian countries obtained freedom scores between 35 and 70 and are classified as partly free. The rest of the world, Central Africa, part of Southeast Asia, Central Asia, and Middle East countries received a freedom score below 35 and are considered not free.

Death Rate and Freedom Score

To explore which variables are the most related to the heart disease death rate, we relate 'Freedom_Score' with the 'Death_Rate'. From the scatter plot, we cannot find an apparent relationship between 'Freedom Score' and 'Death Rate'. Most of the countries, regardless of the freedom score, have the Death Rate in lower ends, clustering around 0.2%. A greater number of countries were observed to have a death rate greater than 0.2% among the countries with a freedom score above 50. The countries with a freedom score greater than 75 have more points above the death rate of 0.2%. However, the points were so dispersed that they were not able to give insights into the trends. The line of the best fit displays a gentle linear slope, almost like a horizontal line, until the death rate is 0.2%. The correlation coefficient was computed to be 0.1816, which suggests that the two variables are very loosely related.

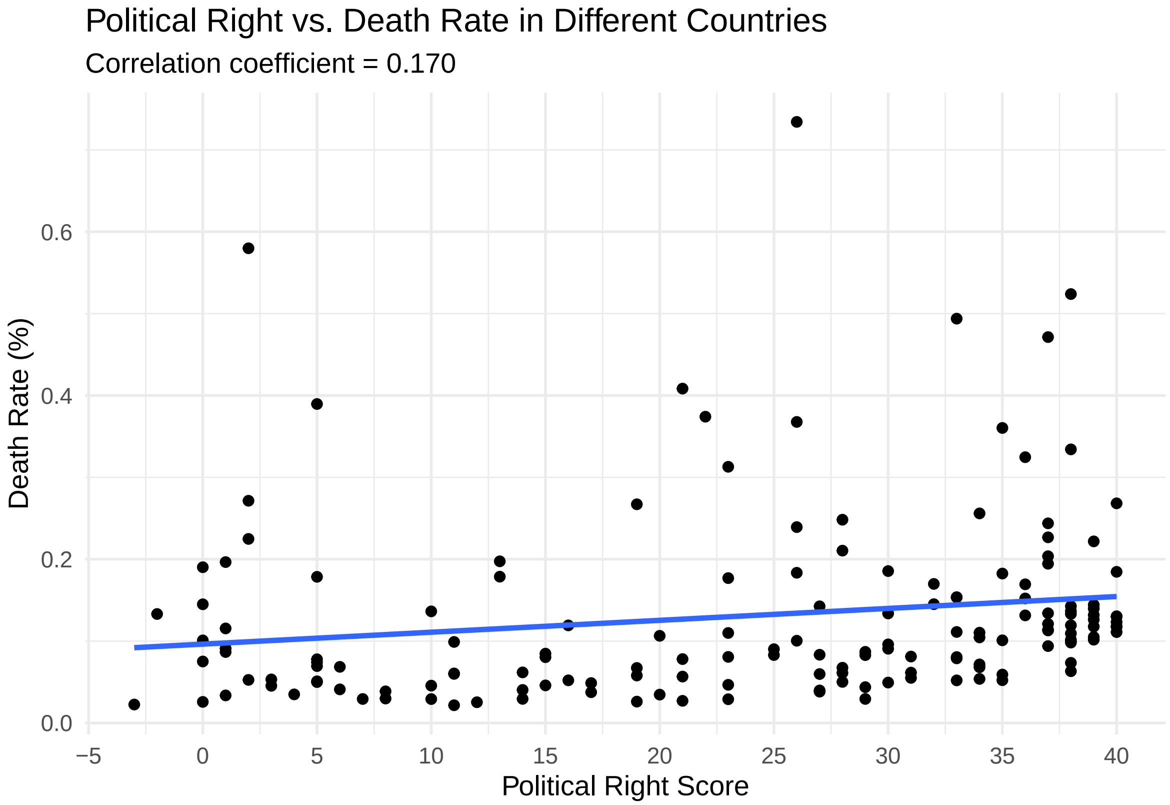


The plot was further separated by freedom status to explore the patterns in death rate related to freedom status. Unfortunately, even with freedom statuses, there were no explainable patterns found. In “Not Free” countries, the death rate demonstrated a slightly decreasing pattern with an increase in freedom score. However, in both “Partly Free” and “Free” countries, the death rate illustrated a slightly increasing pattern with an increase in freedom score. Nonetheless, those patterns are not strong enough to explain the noticeable trend between Freedom and Death Rate.

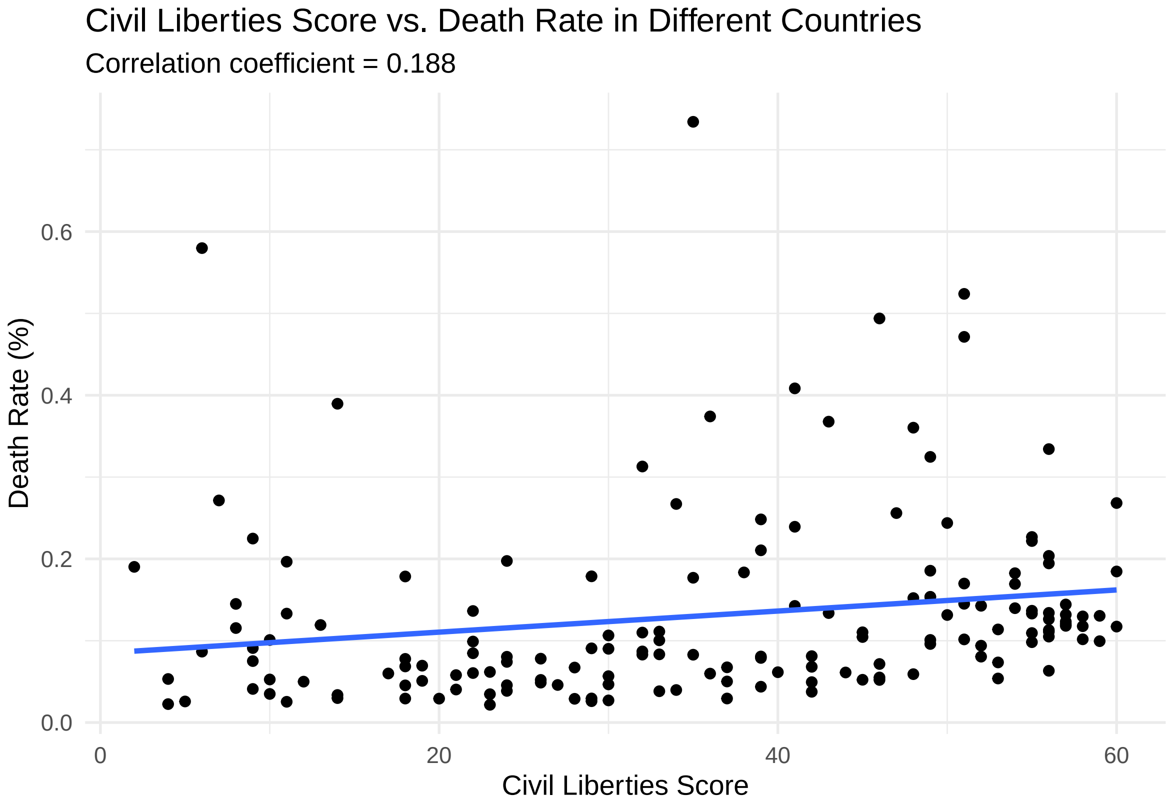


Thorough Look into Political Rights and Civil Liberties

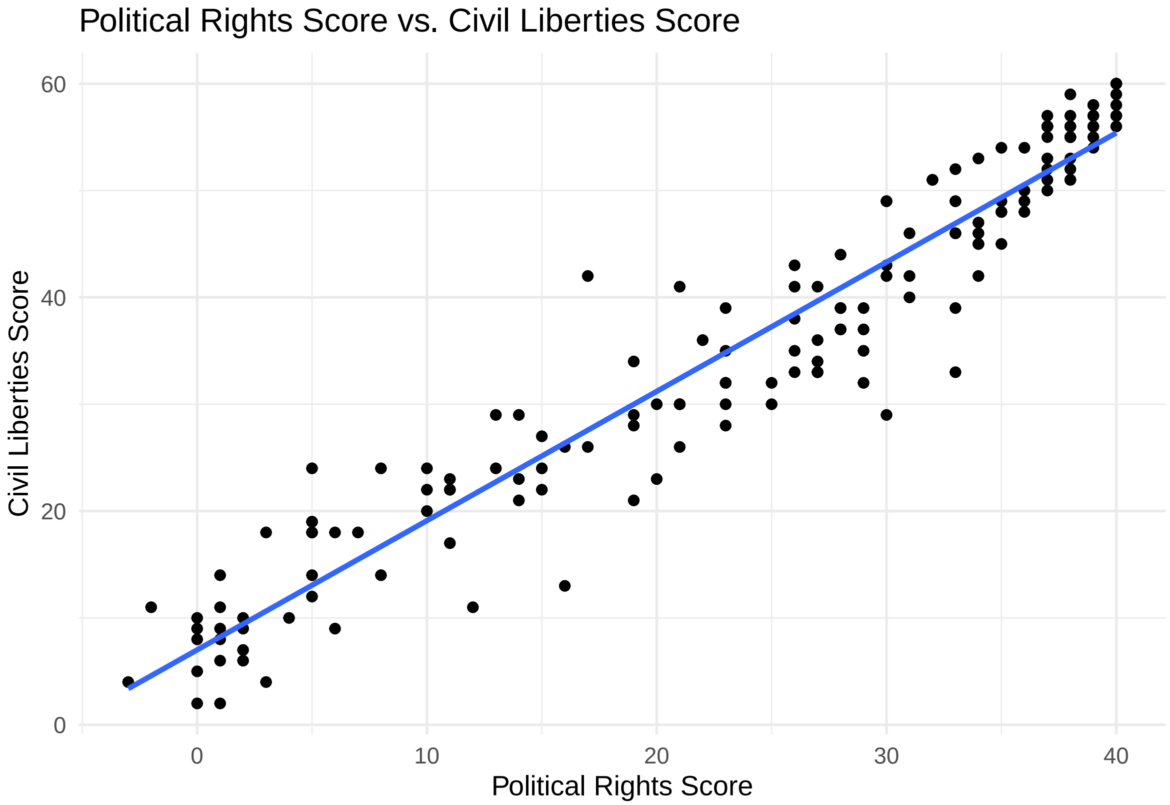
Further analysis was conducted to explore the relationship between political rights, civil liberties, and the death rate. Would political rights and civil liberties align well with each other and thus demonstrate similar patterns when related to the death rate? The intention was to detect anomalies in political rights and civil liberties.



Political rights measure if the chief national authorities and legislative representatives are elected through free and fair elections, if people have the right to organize different political parties, if the opposing party can gain power through a fair election without external forces, and if the government is operated with transparency. Based on the analysis, political rights do not have an apparent relationship with the death rate. The points are all spread out throughout the plot. There is very little correlation observed between political rights and the death rate. One noticeable aspect in the political right section is that some countries have negative scores. These are not errors. In fact, countries like China and Syria received a -3 point deduction for "occupying power deliberately and changing the ethnic composition of a country or territory so as to destroy a culture or tip the political balance in favor of another group". Not gaining many points in other political right sections, those countries end up having negative scores. Overall, the correlation coefficient is calculated to be 0.1694, which is close to that of the Freedom score in the previous analysis.



In civil liberties sections, the countries are scored based on if the media functions as an independent entity, and whether individuals can exercise freedom of religion, speech, and assembly. The countries are evaluated for having an independent judiciary, due process in civil and criminal matters, and equal treatment of the population in front of laws and policies. Lastly, civil liberties include individual rights such as freedom of movement, right to own property and start own business, freedom of marriage, and equal opportunity. Previously in the political right section, it was mentioned that some countries have negative scores, but in civil liberties, no such result was observed. Since all the countries have a greater score for civil liberties score than that of political rights, no country would have Freedom Score in the negative range. From the analysis, there wasn't much noteworthy trend. Although the countries with high civil liberties had a greater number of high death rates, the death rates are randomly plotted in no apparent relation to the civil liberties score. The correlation coefficient is calculated to be 0.188, which is close to that of the Freedom score.



When we analyze the relationship between political rights and civil liberties, the two variables are closely aligned with a correlation coefficient of 0.961. The result suggests political rights scores are connected to the civil liberties score. Countries with high political scores are very likely to have high civil liberties, and countries with low political scores are most likely to have low civil liberties, with no exceptions. From the analysis, we learn that Freedom Score and its subcategories: the political rights and civil liberties scores, do not have a strong relationship to the heart disease rate around the world.

Heart disease and GDP per capita

Definition of GDP

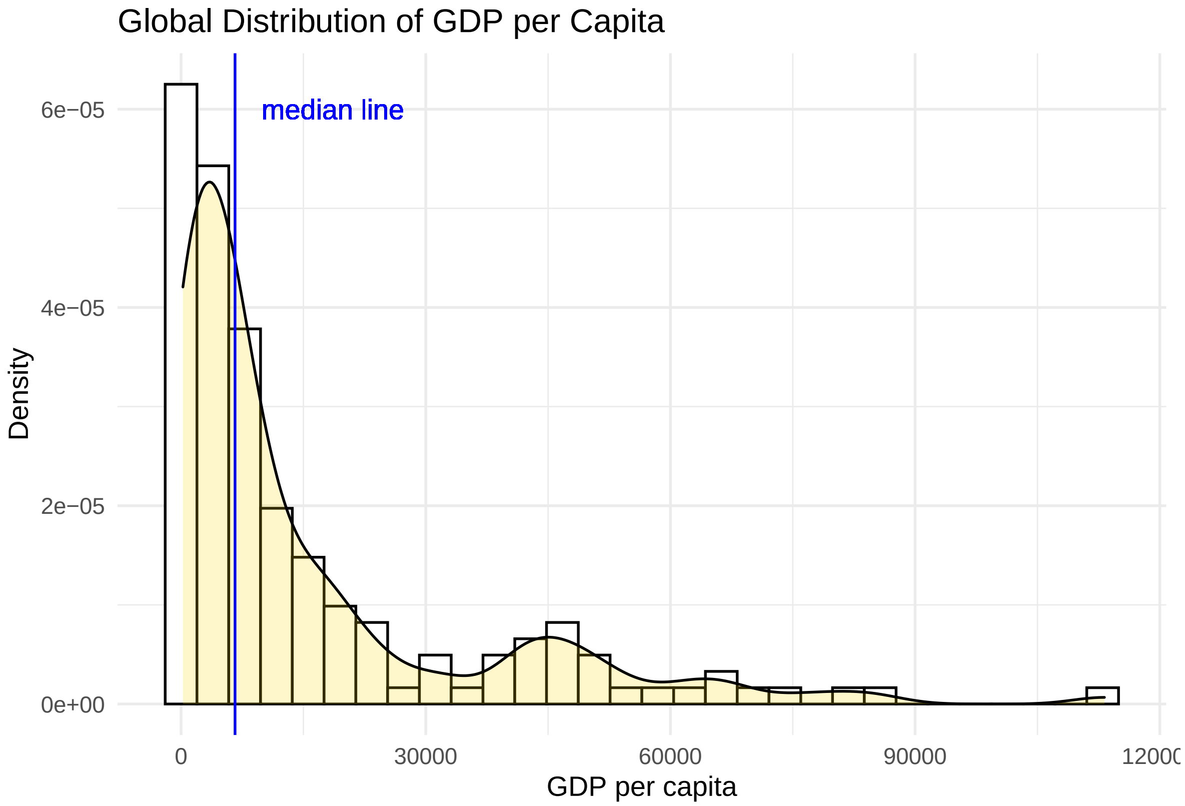
Having looked at political factors, we now turn our attention to economic indicators, starting with Gross Domestic Product (GDP). GDP is one of the most common measures of economic prosperity. It is the final result of the initial distribution of income of all resident units in a country (or region) over a certain period of time (usually one year). It is the total value of final goods and services appropriated by the owners of factors of production in the country in a given period [1].

We assume that GDP might have a negative correlation with death rate because GDP could reflect the living quality of people in this country to some degree. For instance, higher GDP usually accompanies higher average revenue, higher healthcare spending, higher technology and other factors. In this case, we did a correlation analysis between GDP and death rate. Since the absolute value of GDP of each country is significantly different, we use GDP per capita (GDP/population) instead of GDP.

Distribution of GDP per Capita

The distribution of GDP per capita is highly dispersed, the median GDP per capita is around 6617 USD and the 75th percentile is about 17757 USD, which is around 20% of the total range.

From the distribution graph below, the unbalance of world economic development is evident, since the distribution is right-skewed (positive skewness). Due to the high dispersion of the distribution, the correlation between GDP per capita and heart disease death rate might not be strong.



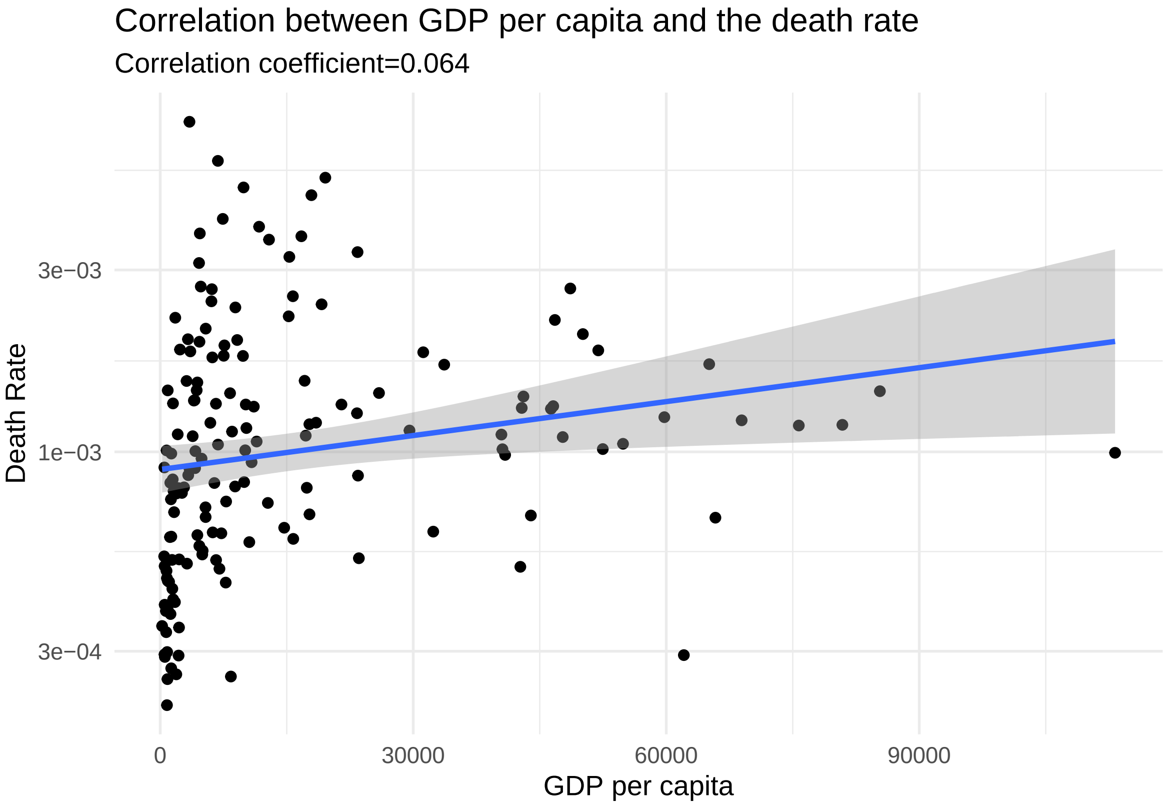
Correalation between GDP per Capita and Heart Disease Death Rate

General correlation

The Correlation coefficient between GDP per capita and heart disease death rate is only about 0.064, which is small and indictates that GDP per capita is irrelevant to heart disease death rate. However, we can find some other information from the graph.

From the correlation graph we notice that most of the countries gather when GDP per capita is lower than 15,000 and those countries are more discrete compared to countries with higher GDP per capita. Considering this, we decided to calculate the correlation coefficient by groups with different values for GDP per capita.

Range of GDP	Number of Countries
0-2,500	44
2,500-5,000	24
5,000-7,500	17
7,500-10,000	13
10,000-15,000	11
15,000-20,000	14
20,000-25,000	5
25,000-30,000	2
30,000-35,000	3
35,000-40,000	0
40,000-45,000	7
45,000-50,000	5
50,000 <	13



Correlation by groups

We divided all the countries into 13 groups according to the GDP per capita:

Since the number of countries whose GDP per capita is between 20,000 and 50,000 are small, we decided to take 20,000 to 35,000 and 40,000 to 50,000 as new groups to be analyzed.

From the table we can see that there are 44 countries with GDP per capita lower than 2,500 (which could be regarded as poor), which is the highest among the countries. There are 98 countries with GDP per capita lower than 10,000, accounting for more than 60% of the all countries in our data set. The number of countries with GDP per capita more than 20,000 is 35 (which could be regarded as close to developed) and there are

Range of GDP per Capita	Correlation Coefficient
0-2500	0.3495849
2,500-5,000	0.04892705
5,000-7,500	0.2945497
7,500-10,000	0.4187217
10,000-15,000	0.1472846
15,000-20,000	0.2134812
20,000-35,000	-0.07328717
40,000-50,000	0.612101
50,000 <	-0.217632

13 countries whose GDP per capita is more than 50,000 (which could be regarded as well developed). From this we further corroborate that the variance of GDP per capita between the countries is high.

From the figure for correlation coefficients by groups, we can find that when GDP per capita is lower than 10,000, the correlation coefficient is generally higher than the correlation coefficient when GDP per capita is higher than 10,000. However, for the 2,500 to 5,000 range, the correlation coefficient is only approximately 0.05, which is too small to indicate a relationship. Also, the other correlation coefficients when GDP per capita is 10,000 approximately range between 0.30 and 0.42, and therefore only show weak correlations. When GDP per capita is higher than 10,000, the correlation coefficients are relatively lower, except when GDP per capita is between 40,000 and 50,000, where the correlation coefficient is around 0.61, which could be regarded as a moderate correlation. However, the 0.61 is an abnormal value because all the other coefficients when GDP per capita is higher than 10,000 are relatively much lower and could be regarded as a weak or very weak correlation.

We do not have enough information to identify the reason why the correlation coefficient is abnormal when GDP per capita is between 40,000 and 50,000. One possibility might be that the number of observations for this range is too small, only 12 countries. In general, the correlation between GDP per capita and heart disease death rate shows a weak positive correlation : higher GDP per capita relates weakly to higher death rates from heart disease.

This result conflicts our original assumption, sine we expected a negative relationship between GDP per capita and heart disease death rate.

To analyze this issue further, we decided to conduct some analysis using a different economic indicator to see if this could shed more light on the relationship between economic prosperity and heart disease death rate.

Heart Disease and HDI

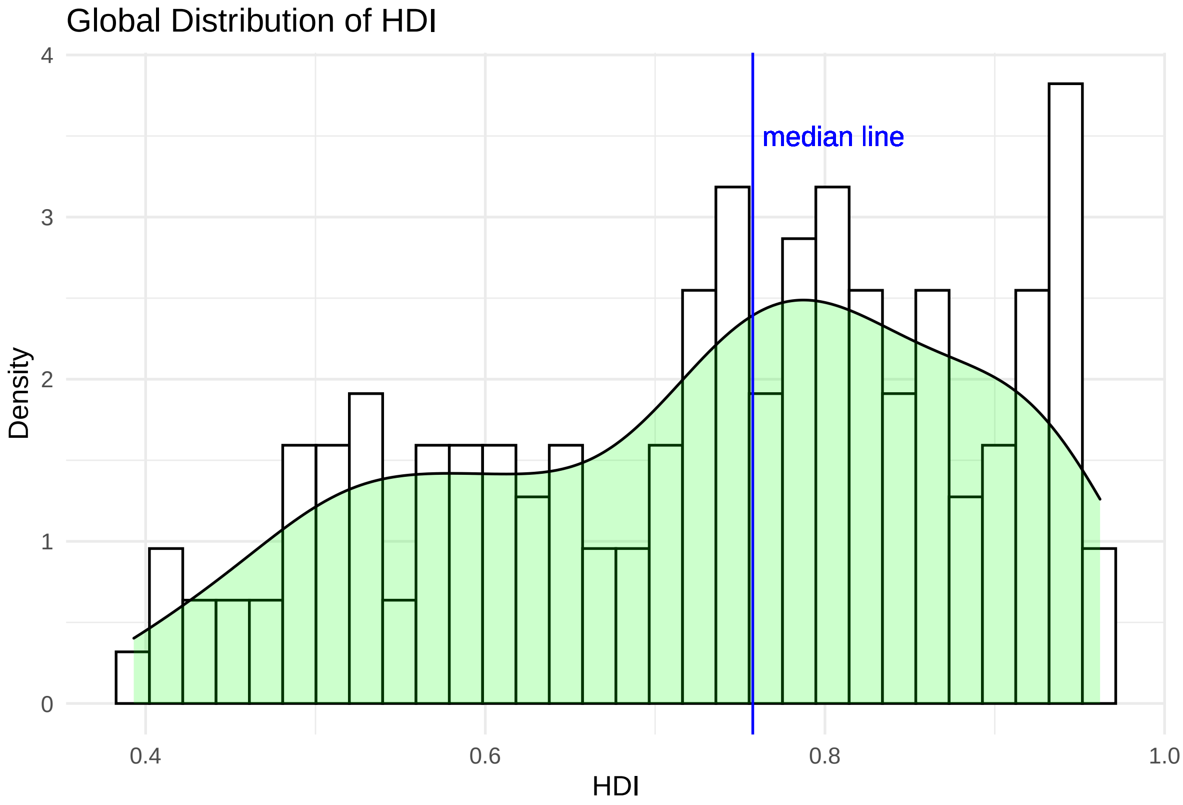
Definition of HDI

The Human Development Index (HDI) is an indicator proposed by the United Nations Development Programme (UNDP) in the Human Development Report 1990 to measure the level of economic and social development of the United Nations member countries, as a result of the challenge to the traditional GNP indicators.

HDI is a composite indicator based on three basic variables: life expectancy, education level and quality of life. Considering this, HDI might be a better indicator of people's living conditions compared to GDP per capita. Since 1990, the Human Development Indicators have played an extremely important role in guiding developing countries to formulate appropriate development strategies. Since then, the United Nations Development Programme has published the HDI for each country in the world every year and used it in the Human Development Report to measure the level of human development in individual countries [2].

Distribution of HDI

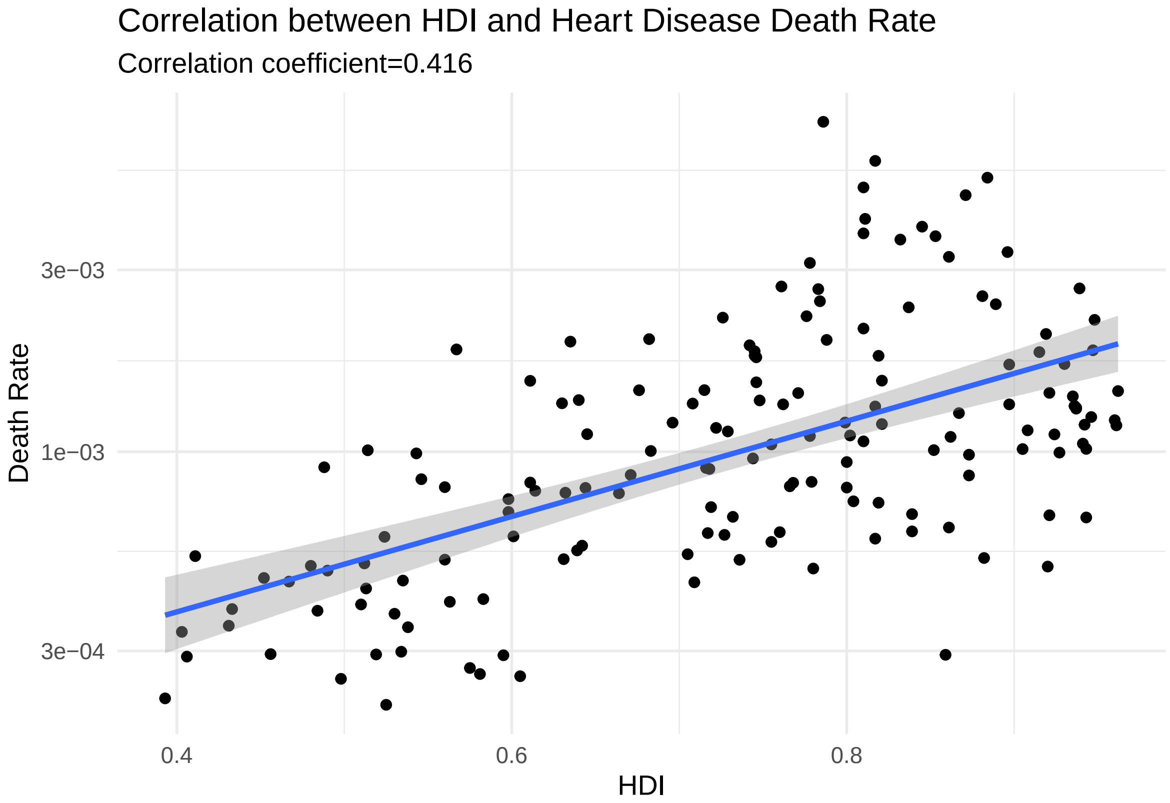
Compared to the distribution of GDP per capita, HDI has a more uniform distribution. The variance is lower and the median line is closer to the middle of the x-axis. This suggests that the world is more equal in terms of HDI when compared to GDP per capita.



Correlation between HDI and Heart Disease Death Rate

Similarly to GDP per capita, we first assume that HDI has a negative correlation with heart disease rate because HDI ccounts for factors such as life expectancy and quality of life. With higher living quality, the heart disease death rate might be lower.

However, our analyis shows that the correlation coefficient is still positive and the value is 0.416, which is a low positive correlation. From the correlation plot below, we find that the points are evenly distributed along the X-axis. For a certain HDI level, we also find that the death rate ranges a lot. This might be because heart disease death rate is a complex issue and is thus related to many other factors that are not accounted for in this analysis.



Summary of Economic Indicators and Heart Disease Death Rate: An Unexpected Finding

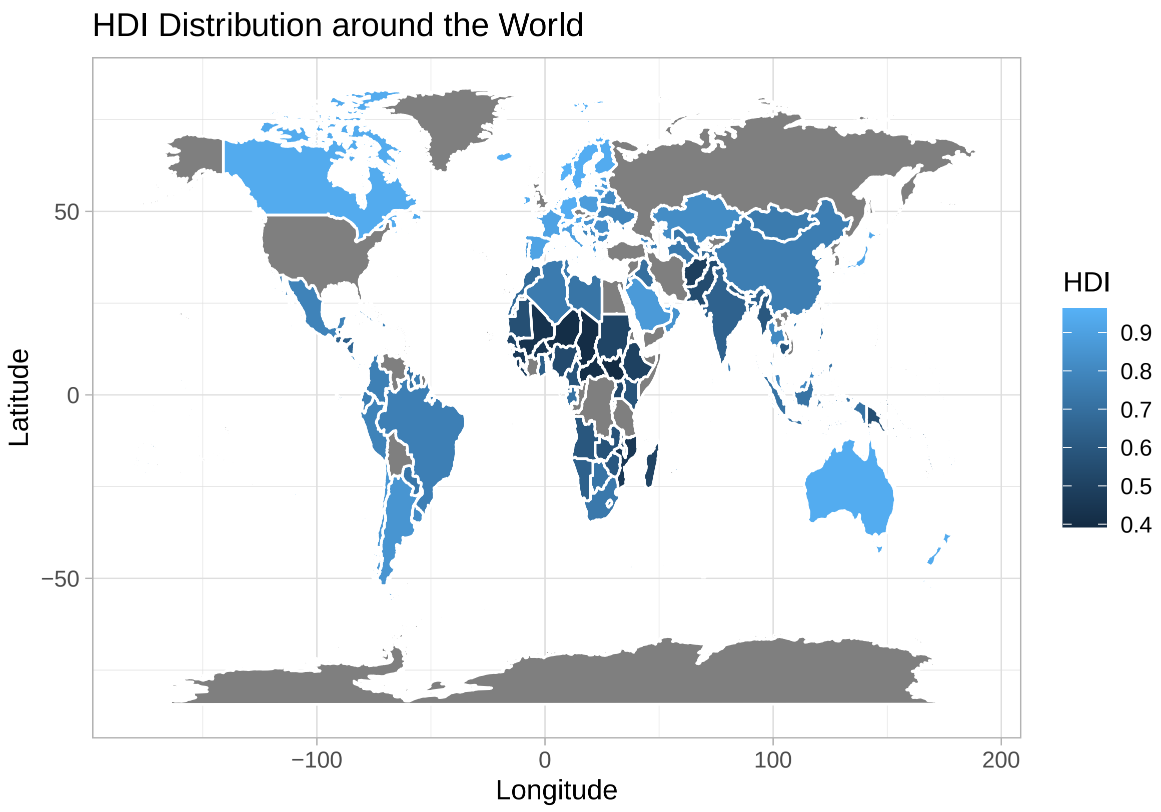
Our analysis has shown that both GDP per capita and HDI index show a low positive correlation with heart disease death rate. This is a non-intuitive finding. Although the strength of this correlation is only weak, we have not found any evidence for a negative correlation between these variables. This is surprising since it would be reasonable to assume that countries with greater economic prosperity and a higher quality of life should have lower rates of heart disease death rate. However, our analysis shows that this does not appear to be true. For a possible answer, we consult work done by other researchers

HDI, Climate, and Heart Disease Death Rate

At this point in our study we have looked at heart disease from several key standpoints, including economic indicators, climate, and political factors. Thus, at this stage, we wish to put some of these together to develop a holistic view of heart disease around the world and the factors that are related to it.

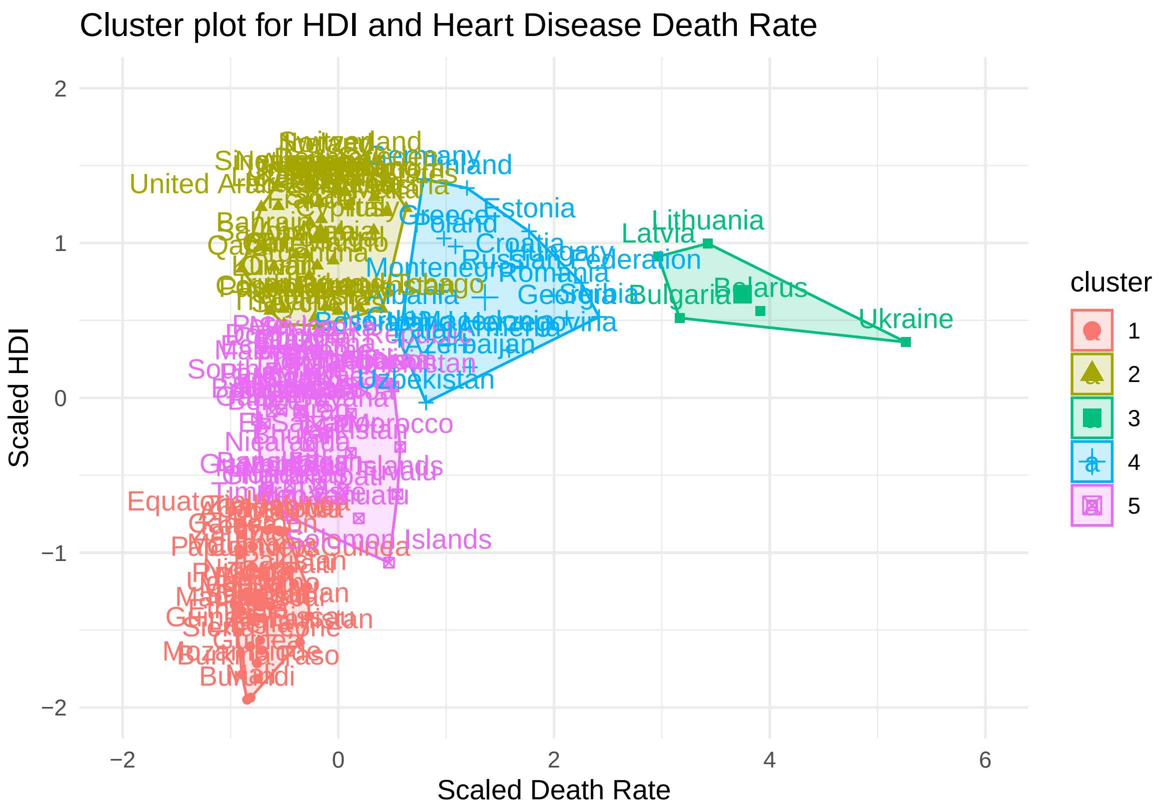
HDI distribution around the world

We will commence this section of the report by first visualizing the distribution of HDI around the world. This will provide us with a good context in which to conduct the remainder of the analysis.

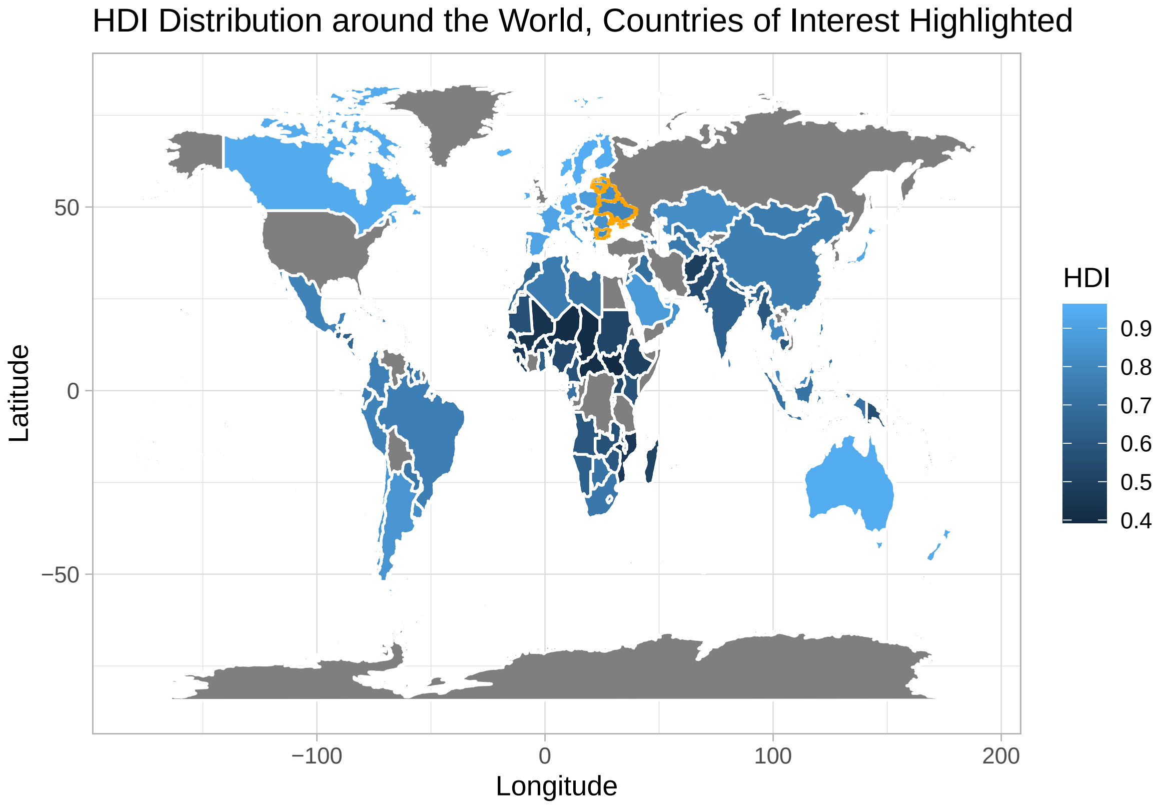


This world map is colored by different levels of HDI. The lighter the color means the larger the value of HDI. From this map, we can observe that in Canada, Australia, and most countries of Europe, HDI values are much higher than others.

We have already established that HDI is weakly positively correlated with heart disease death rate. We also saw that Eastern European countries show higher rates of heart disease than in other regions. Therefore, we would like to examine if these three different factors coincide together in some way. To accomplish this, we start by conducting a second K-Means clustering analysis, but this time the variables we use for our clustering algorithm are HDI and heart disease death rate.



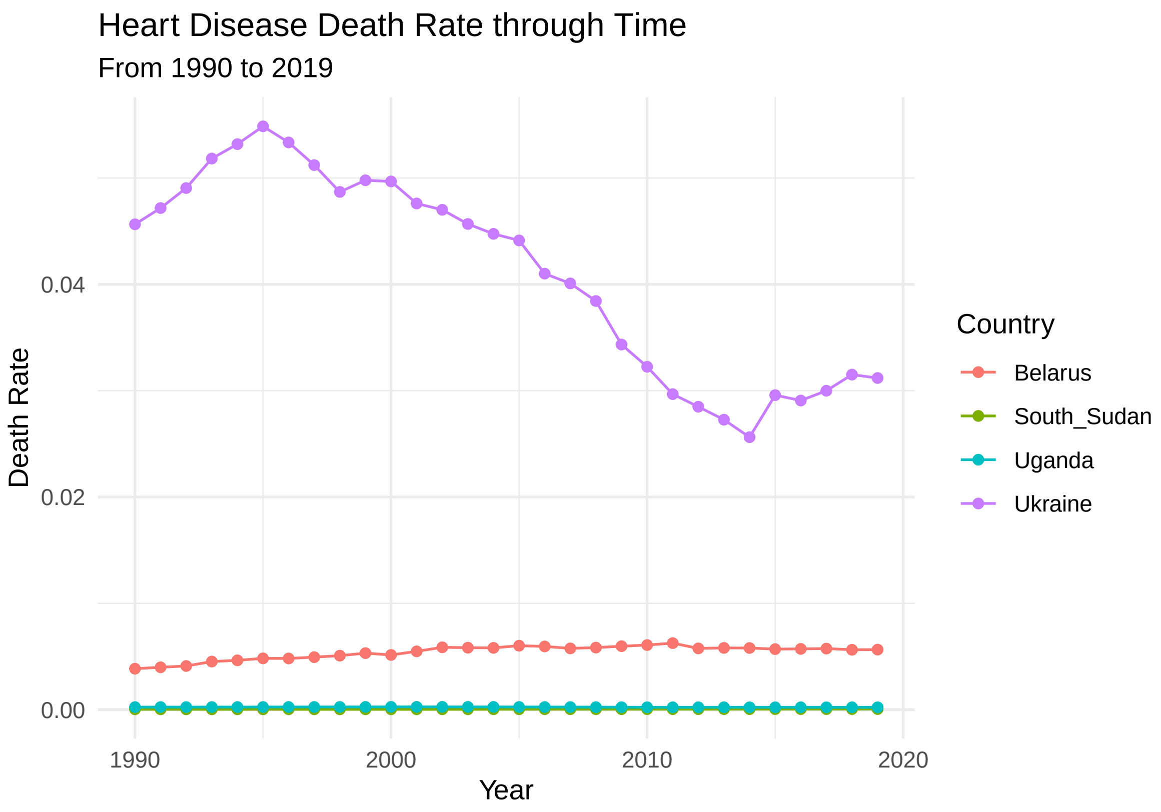
From this cluster plot, we once again find that five Eastern European countries, Lithuania, Latvia, Bulgaria, Belarus, and Ukraine, are clustered together. This indicates that they have a similar HDI. In fact, the majority of these countries are ranked highly in terms of HDI. To summarize our findings in a single visualization, we select these five countries to generate the map visualization below.


HDI Distribution around the World, Countries of Interest Highlighted
Latitude
50
0
−50
Longitude
−100
0
100
200
HDI
0.9
0.8
0.7
0.6
0.5
0.4

In this map, the area with the orange highlight depicts the locations of the Eastern European countries. From this plot, it is evident that these countries are not only in close proximity to each other in terms of geography, but also in terms of their HDI. This ties in with our previous finding that these five countries have the same climate zone, DFB (cold continental without a dry season). Thus, these countries all have a similar climate, HDI, and heart disease death rate. Although it is possible that this finding is due to some unseen variable(s) that have not appeared in this study, this observation does nevertheless provide some support to the theory that deaths from heart disease are in some way related to the climate and level of economic development of the country.

Phase 3: Time Series Analysis

We have successfully identified the countries that suffer from the highest rates of death due to heart disease. To conclude our analysis, we would like to take a look through time to see how some of these countries arrived at this point. The central question we wish to address is whether heart disease death rate has always been so high for these countries. Additionally, we want to use historical data to determine the most recent trends in heart disease death rates, which may enable us to make some predictions about how this could change in the future.



We decided to plot heart disease death rate against time for two of the countries that currently have some of the highest heart disease death rates, Belarus and Ukraine, as well as two of the countries that currently have some of the lowest heart disease death rates, South Sudan and Uganda. From our resulting time series plot, we see that the heart disease death rate for Uganda and South Sudan has remained essentially constant at close to zero from 1990 to 2019. The heart disease death rate for Belarus has also remained fairly constant. The country that has experienced the most change in heart disease death rate over these three decades is Ukraine. Based on this plot, we can see that heart disease death rate in Ukraine peaked in the year 1995, after which it fell rapidly until 2014. After 2014, there was a slight increase in the heart disease death rate, but it appears to have started to plateau in more recent years. Thus, this time series analysis suggests that heart disease mortality rates are generally projected to remain fairly constant in the near future.

Conclusions

In this study, we presented some insightful findings about death rate due to heart disease around the world and the factors that may be related to it. We first provided evidence that heart disease was the disease that killed the most people globally as per our most recent data (2019). This emphasizes the relevance and necessity to better understand factors that contribute to heart disease death rates. We then examined this issue from multiple perspectives, allowing us to determine that the countries that presently suffer from the highest rates of heart disease mortality are economically well-developed and concentrated in the Eastern European region. These countries share a similar level of HDI and similar climates (predominantly a cold continental climate). Therefore, although we did not find a strong relationship between heart disease and these variables in isolation, we did find some evidence to suggest that there may be a relationship between deaths from heart disease and specific combinations of economic development and climate, although the precise reasons behind this relationship are beyond the scope of this study. Finally, we looked at how heart disease death rate has changed over the 30 year period from 1990 to 2019 for a select few countries, allowing us to predict that global heart disease death rates will generally remain fairly constant for the foreseeable future, based on trends in recent years.

Limitations and Future Work

In this report, we have presented a detailed study of heart disease death rate on a global scale and the factors that may be related to it. However, this work is nonetheless not without flaws. One key limitation is that we were unable to find reliable data for all the countries in the world. After combining data from multiple sources to obtain information for all variables of interest, we only had complete data for 160 countries for the main portion of our study. This is approximately 82% of all countries. However, 18% is not an insignificant percentage, and thus it is not improbable that we may be missing out on important insights by not including these remaining countries in our study. Thus, one possible avenue for future work is to conduct a similar project to the study that we have presented here that includes data from these countries. This will help to corroborate our findings and may also lead to new discoveries. The second key limitation of this study is that, although we have identified certain patterns in heart disease mortality rates around the world, we do not have enough information to deduce the reasons why these patterns exist. The most significant finding of our project is that the countries that presently suffer from the highest rates of death due to heart disease are all located in Eastern Europe and share a similar climate and level of economic development. Considering their geographic proximity, we hypothesize that this finding might actually result from factors such as diet and culture, since countries that are neighbors and of similar economic development tend to have a lot in common in these areas. However, our data is insufficient to investigate this hypothesis. Considering this, future work can build upon our study by focusing on just this group of countries. By collecting and analyzing the relevant data, future work could test this hypothesis.

References

[1] OECD (2022), OECD Economic Outlook, Volume 2022 Issue 2: Preliminary version, OECD Publishing, Paris, https://doi.org/10.1787/f6da2159-en.

[2] Human Development Reports, 2019.HUMAN DEVELOPMENT REPORT 2018-19. Publisher name. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

[3] Human Development Index (HDI) by Country 2022. https://worldpopulationreview.com/country-rankings/hdi-by-country

[4] Weather and Climate. (2022), World Climate Data. List of countries by climate zone and average yearly temperatures. Retrieved December 7, 2022, from https://tcktcktck.org/countries

[5] Arnfield, J. (2022), Köppen Climate Classification. Encyclopædia Britannica. Retrieved December 6, 2022, from https://www.britannica.com/science/Koppen-climate-classification

[6] World Health Organization. (2020, December 9). The top 10 causes of death. World Health Organization. Retrieved December 6, 2022, from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[7] Freedom House. (2020), Countries and territories. Freedom House. Retrieved December 7, 2022, from https://freedomhouse.org/countries/freedom-world/scores

[8] Belarus population 1950-2022. MacroTrends. (2022), Retrieved December 7, 2022, from https://www.macrotrends.net/countries/BLR/belarus/population

[9] South Sudan population 1950-2022. MacroTrends. (2022), Retrieved December 7, 2022, from https://www.macrotrends.net/countries/SSD/south-sudan/population

[10] Uganda population 1950-2022. MacroTrends. (2022), Retrieved December 7, 2022, from https://www.macrotrends.net/countries/UGA/uganda/population

[11] Ukraine population 1950-2022. MacroTrends. (2022), Retrieved December 7, 2022, from https://www.macrotrends.net/countries/UKR/ukraine/population

[12] Institute for Health Metrics and Evaluation. (2022, December 6). Retrieved December 8, 2022, from https://www.healthdata.org/