**Acknowledgments­­­**

I would like to express my profound gratitude to Dr. Huynh De Thu, an esteemed lecturer specializing in Data Analysis and Visualization with Python. Dr. Thu generously shared his time and profound expertise, providing invaluable guidance throughout the completion of this paper. His mentorship not only served as a strong source of motivation but also inspired me to strive for excellence and achieve the best possible outcomes.

I also extend my heartfelt thanks to the team at Our World in Data, particularly Dr. Saloni Dattani, Dr. Fiona Spooner, Professor Max Roser, and Dr. Hannah Ritchie, eminent researchers in the fields of global health and data analysis. Their contributions not only provided invaluable data but also imparted extensive knowledge and insights, deepening my understanding of community health issues and global development. Their research endeavors have significantly facilitated and laid a solid foundation for this paper.

Contents

[I. Introduction 4](#_Toc163853903)

[I.1 Purpose of the paper 4](#_Toc163853904)

[I.2 Introduction to the dataset 4](#_Toc163853905)

[I.3 Introducing some of the libraries and configurations for Python used in this paper 6](#_Toc163853906)

[II. Data Exploration 7](#_Toc163853907)

[II.1. Dataset description 7](#_Toc163853908)

[II.2. Dataset description 8](#_Toc163853909)

[III. Data preprocessing 11](#_Toc163853910)

[III.1. Read and process redundant data 11](#_Toc163853911)

[III.2. Data classification 11](#_Toc163853912)

[IV. Data Analysis and Visualization 14](#_Toc163853913)

[IV.1. Overall trend analysis and data visualization 14](#_Toc163853914)

[V. Conclusion 20](#_Toc163853915)

[VI. References 20](#_Toc163853916)

1. **Introduction**
   1. Purpose of the thesis
   2. Introduction to the dataset
2. **Data Exploration**
   1. Data description
   2. Data preprocessings
3. **Data Analysis and Visualization**
   1. Overall trend analysis and data visualization
   2. Comparison between countries and data visualization
   3. Analysis of deaths due to infectious diseases and data visualization
   4. Analysis of deaths due to traffic accidents and data visualization
   5. Analysis of deaths due to postpartum disorders and data visualization
   6. Analysis of deaths due to cardiovascular diseases and data visualization
   7. Analysis of deaths due to lower respiratory diseases and data visualization
   8. Analysis of deaths due to suicide and data visualization
   9. Analysis of deaths due to diabetes and data visualization
   10. Analysis of deaths due to chronic kidney disease and data visualization
   11. Specific analysis for Vietnam and data visualization
4. **Conclusion**
   1. Summary of analysis results
   2. Proposal for further research directions
5. **References**

# Introduction

## I.1 Purpose of the paper

The purpose of this paper is to analyze and visualize data on the causes of death across different countries and years, using Python as the primary tool. This analysis aims to answer the critical question: “What are people dying from?”

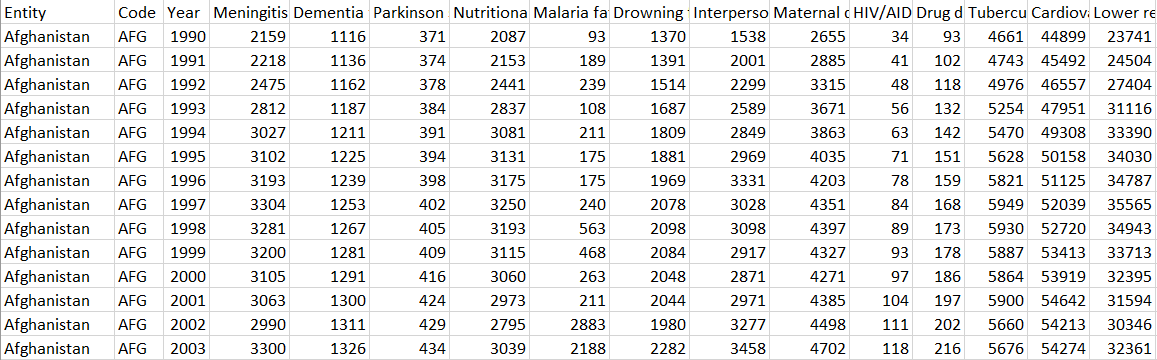
Understanding the causes of death is essential for guiding decisions in public health and finding ways to save lives. Many leading causes of death receive little mainstream attention, and this analysis aims to bring these causes to light.

This paper will explore how death rates from various causes have changed over time and how these changes have led to shifts in the leading causes of death. For instance, while infectious diseases once dominated, death rates from these diseases have fallen quickly, leading to non-communicable diseases such as heart diseases and cancers becoming the most common causes of death globally.

Furthermore, this paper will delve into how this data can help understand the burden of disease more broadly and offer a lens to see the impacts of healthcare and medicine, habits and behaviors, environmental factors, health infrastructure, and more.

By providing a comprehensive analysis and visualization of this data, this paper aims to contribute to the ongoing global conversation about public health and disease prevention. Ultimately, the goal is to use this analysis to inform strategies that can further reduce the impact of causes of death and improve health outcomes worldwide.

## I.2 Introduction to the dataset

The dataset utilized in this paper is a CSV file compiled from 210 countries and 18 international organizations spanning the years 1990 to 2019.

*Table 1.1: A piece of data from the dataset*

Dr. Saloni Dattani, a distinguished researcher, authored this dataset in collaboration with renowned experts in the field of global health. The dataset can be accessed and downloaded from [Our World In Data](https://ourworldindata.org/causes-of-death), a prominent platform dedicated to providing comprehensive data-driven insights into global issues.

Dr. Saloni Dattani, a researcher specializing in health-related topics, particularly mental health, joined the project team in 2021. Her background includes a Ph.D. in psychiatric genetics from the University of Hong Kong and King’s College London.

Additionally, the project team includes esteemed professionals contributing to the field of data science and global development:

Dr. Fiona Spooner: A Senior Data Scientist who joined the team in 2021. Dr. Spooner's expertise lies in tracking Sustainable Development Goals and modeling the COVID-19 pandemic. She holds a Ph.D. in Ecology and Environment from UCL (London, UK) and an MSc in Conservation Science from Imperial College (London, UK).Professor Max Roser: The Founder and Executive Co-Director of Our World in Data. Professor Roser initiated this influential publication in 2011 and continues to lead its development. He holds the position of Professor of Practice in Global Data Analytics at the University of Oxford’s Blavatnik School of Government. Additionally, he serves as the Programme Director of the Oxford Martin Programme on Global Development and Executive Co-Director of Global Change Data Lab.

Dr. Hannah Ritchie: Serving as the Deputy Editor and Science Outreach Lead at Our World In Data since 2023, Dr. Ritchie is an integral member of the team. Her focus areas include long-term development in food supply, agriculture, energy, and the environment. Dr. Ritchie earned her Ph.D. in GeoSciences from the University of Edinburgh and has made significant contributions to research as the former Head of Research.

The dataset provides a comprehensive overview of causes of death globally, encompassing various categories from infectious diseases to non-communicable diseases and external causes like road injuries and suicide. Each entry in the dataset represents a country in a specific year, with columns detailing different causes of death and corresponding values indicating the number of deaths attributed to each cause. This dataset serves as a crucial resource for understanding global health trends and evaluating the efficacy of public health interventions. Through analysis, it enables insights into the changing landscape of leading causes of death over time and across different regions, facilitating informed decision-making and targeted interventions in disease prevention and health promotion efforts.

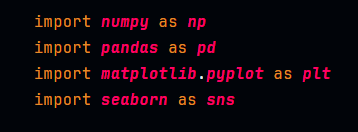
## I.3 Introducing some of the libraries and configurations for Python used in this paper

In this paper, I leverage several Python libraries to analyze and visualize the dataset effectively. The following libraries are utilized:

**NumPy (np)**: NumPy is a fundamental package for scientific computing in Python. It provides support for arrays, matrices, and mathematical functions, making it essential for numerical computations and data manipulation.

**Pandas (pd):** Pandas is a powerful library for data manipulation and analysis. It offers data structures like DataFrame and Series, along with functions to clean, transform, and analyze data efficiently. Pandas is particularly useful for handling structured data, such as the dataset under examination.

**Matplotlib (plt):** Matplotlib is a versatile library for creating static, interactive, and animated visualizations in Python. It offers a wide range of plotting functions and customization options, enabling the creation of various types of plots, including line plots, scatter plots, histograms, and more.

**Seaborn (sns):** Seaborn is a statistical data visualization library built on top of Matplotlib. It provides high-level functions for creating informative and attractive statistical graphics. Seaborn simplifies the process of generating complex visualizations by offering built-in themes, color palettes, and statistical estimation functions.

*Table 1.2: Library import syntax in the script.*

Additionally, specific configurations and settings may be applied to enhance the readability and aesthetics of the visualizations produced in this paper. These configurations may include adjustments to plot sizes, font styles, axis labels, and other visual elements to ensure clarity and coherence in the presentation of results.

By leveraging these libraries and configurations, I aim to conduct a thorough analysis of the dataset and present the findings through clear and insightful visualizations, facilitating a comprehensive understanding of global causes of death and their implications for public health policies and interventions

# Data Exploration

## II.1. Dataset description

The dataset used in this essay includes information on the number of deaths from different causes. This data is collected from various sources and compiled in a CSV file.

Specifically, the dataset includes the following columns:

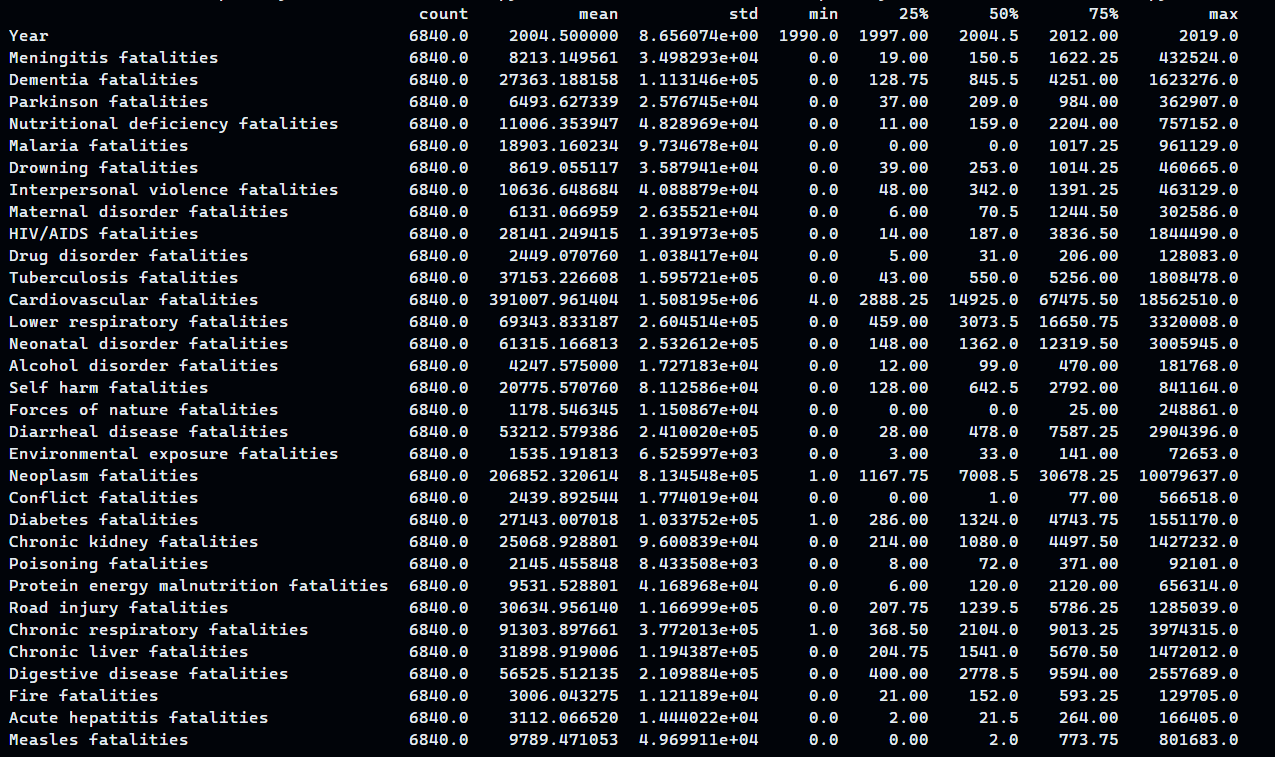
**file\_path = "dataset.csv"  
df = ***pd***.***read\_csv***(file\_path)  
df.***info***()

*Table 2.1: Name of columns in its dataset and data type*

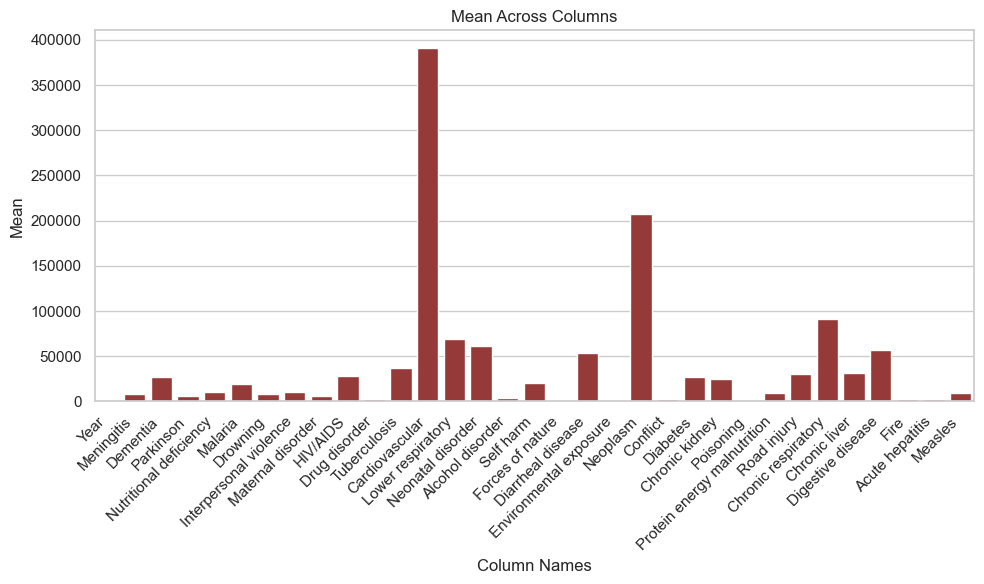
The data set has 35 columns, 6840 rows, and no null values. There are 2 columns with data type String and the remaining 33 columns are int64.

## II.2. Dataset description

Descriptive statistics play a crucial role in analyzing and summarizing the characteristics of a dataset. In this section, we will employ various descriptive statistical measures to gain insights into the dataset on global causes of death.

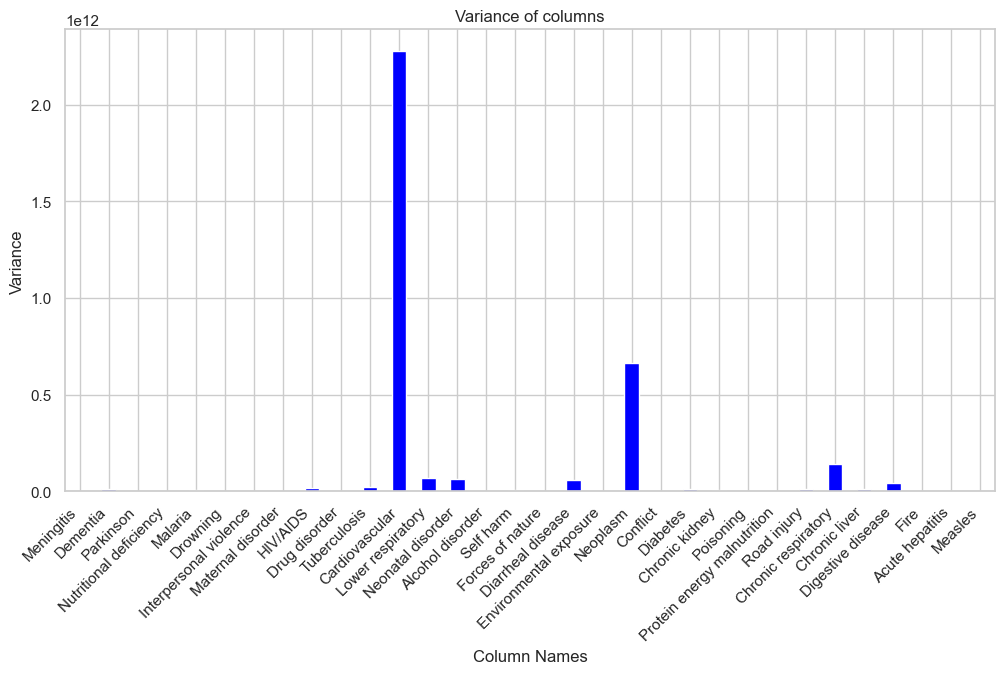
***print***(df.***describe***().T)

*Table 2.2: Detailed descriptive statistics about the columns in the dataset*

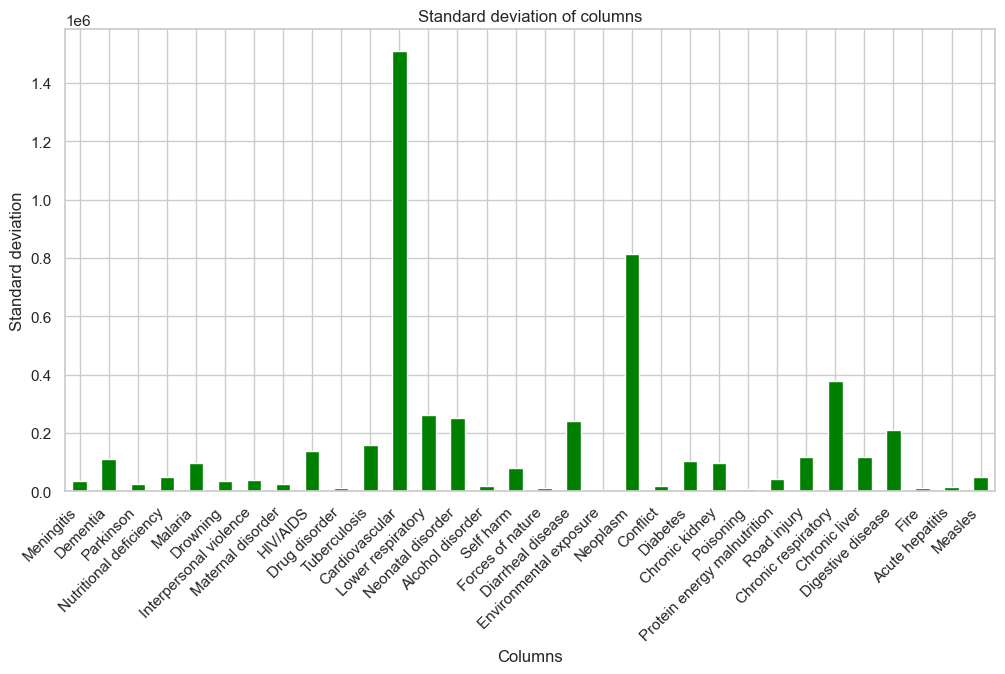


*Table 2.3: The chart means on columns*

Based on the chart, we can see that there are three causes of death with significant average value than other causes, namely "Cardiovascular Fatalities", "Neoplasm Fatalities", and "Chronic Respiratory Fatalities ". This shows that these causes may have caused a large number of deaths during the time and location that the data is collected.



*Table 2.4: The chart variance on columns*



*Table 2.5: The chart standard deviation on columns*

# Data preprocessing

## III.1. Read and process redundant data

After surveying the data set, we can see that the word "fatalities" is repeated too many times in the column names, so I removed it to facilitate information retrieval, information exploitation, and analysis, visualize the data later.

*# Remove the word "fatalities" repeated multiple times in column names to compact the dataframe*

df.columns = df.columns.str.***replace***(' fatalities', '')

## III.2. Data classification

By analyzing the descriptive statistics of the data set in the step above, I created a new data set consisting of columns containing only numeric data types to facilitate statistics and calculations from which to visualize data. Get the data in the best possible way.

*# Remove the 'Year' column and select only numeric columns from the DataFrame*

df\_numeric = df.***drop***(['Year'], ***axis***=1).***select\_dtypes***(***include***=[***np***.number])

Next, I check to see what countries are in the "Entity" column. I realized that, in addition to countries around the world, there are also international organizations that was collected data.

countries = df['Entity'].***unique***()

***print***("Countries included in the dataset:")

for country in countries:

***print***(country)

The "Entity" column has 228 unique values, there are 12 international organizations and 216 countries. Because of concerns that during the calculation process, there may be duplicate data between international organizations and countries.

international\_organizations = ["East Asia & Pacific wb",

                               "Eastern Mediterranean Region who",

                               "Europe & Central Asia wb",

                               "European Region who",

                               "Latin America & Caribbean wb",

                               "G20",

                               "Middle East & North Africa wb",

                               "North America wb",

                               "South Asia wb",

                               "SouthEast Asia Region who",

                               "SubSaharan Africa wb",

                               "Western Pacific Region who"]

df\_international\_organizations = df[df['Entity']/

.***isin***(international\_organizations)]

df\_countries = df[~df['Entity'].***isin***(international\_organizations)]

I separated this data set into 2 parts: **"df\_international\_organizations"** and **"df\_countries".** This will facilitate the analysis and visualization process for future research.

By drawing a pie chart, we can see the proportion between the number of countries and international organizations in the dataset.

*# Plot*

***plt***.***pie***([num\_organizations, num\_countries],

***explode***=(0.1, 0),

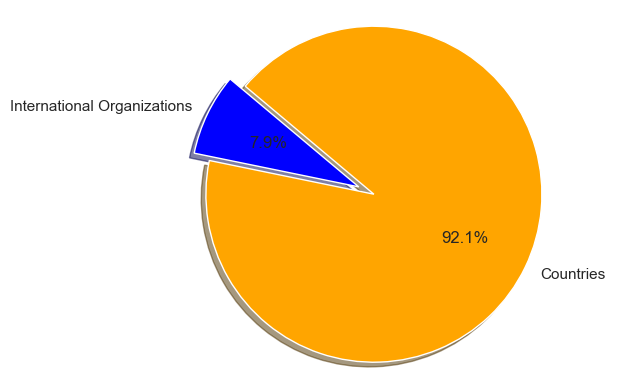
***labels***=('International Organizations', 'Countries'),

***colors***=['blue', 'orange'],

***autopct***='**%1.1f%%**', ***shadow***=True, ***startangle***=140)

***plt***.***axis***('equal')

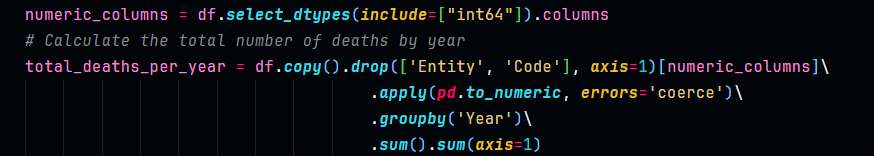
***plt***.***show***()

**

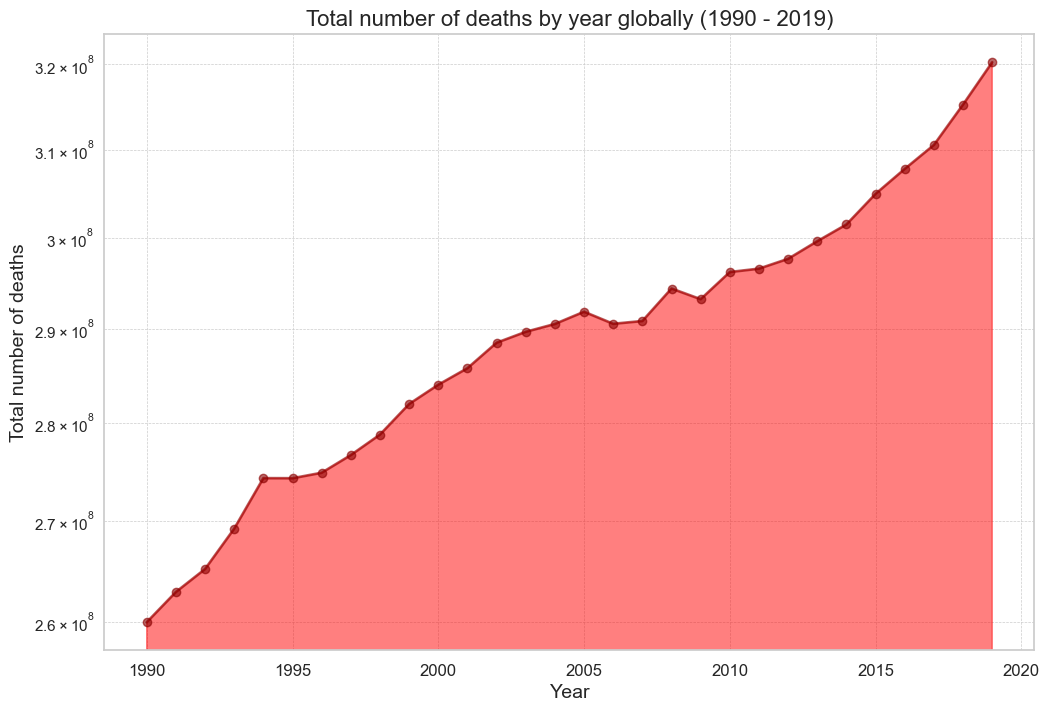
*Table 3.1: The chart shows the ratio of international and   
national organizations in the world in the dataset*

# Data Analysis and Visualization

## IV.1. Overall trend analysis and data visualization

The number of deaths in each region is always different and changes over time. To have a clearer view of the number of deaths worldwide over the years, we can draw a chart for a more intuitive view.

*Table 4.1: Code for visualizing data about the total of deaths   
by year globally in the script.*



*Table 4.2: The chart shows the total number of deaths   
worldwide from 1990 to 2019*

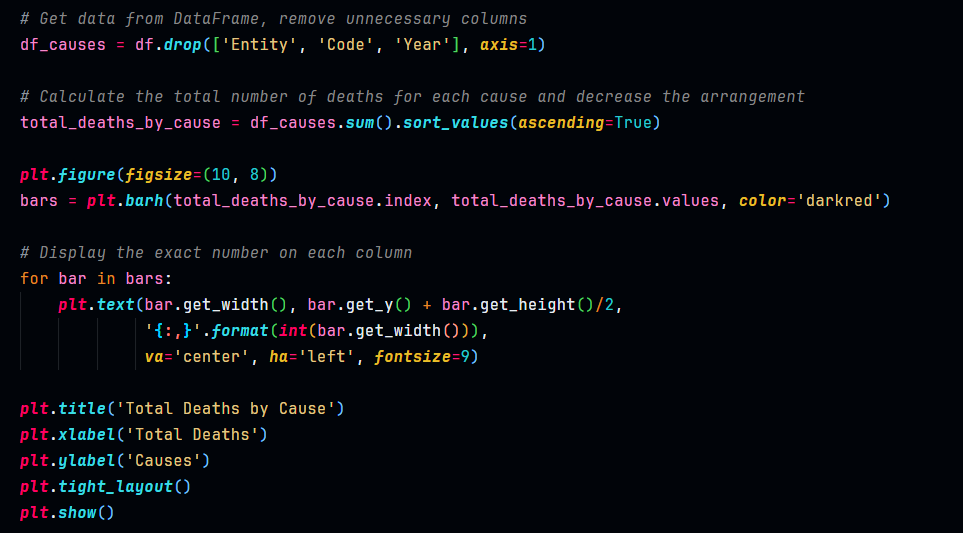
Based on the chart, we can see that the total number of deaths globally has increased very rapidly from 1990 to 2019, from 2.6 billion people to 3.2 billion people in just 30 years. But the increase is not uniform, specifically the number of deaths globally increased rapidly from 1990 to 2005, then it remained stable and increased sharply again in the following years.

The rapid increase in total global deaths from 1990 to 2019 can be explained by several factors:

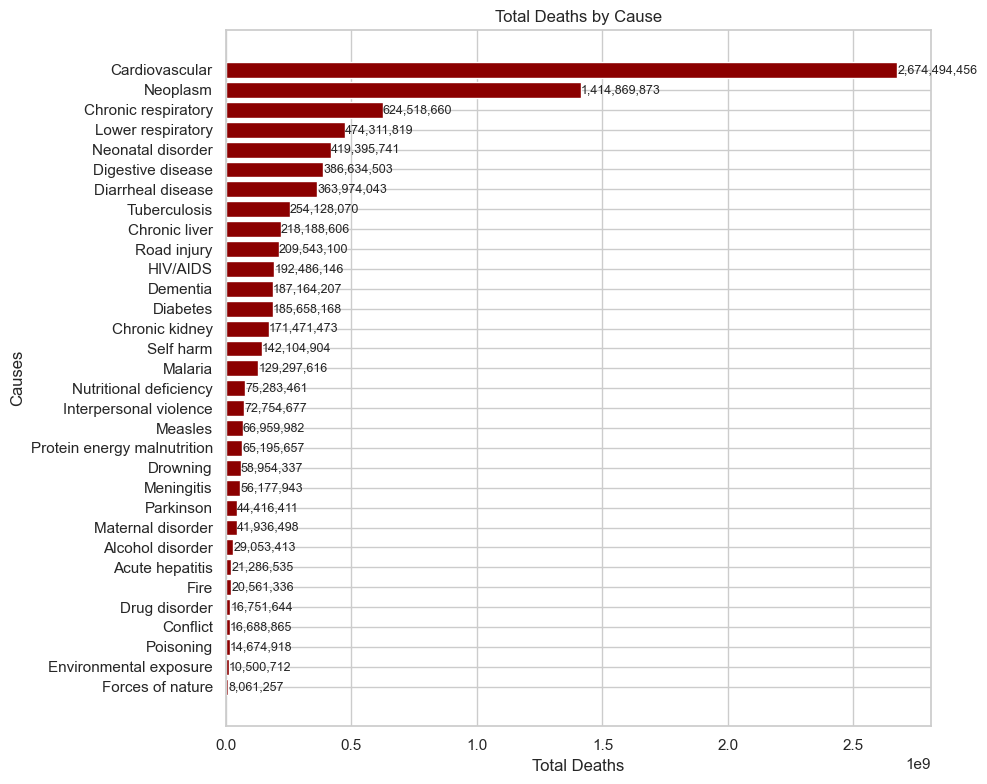
1. Population Growth: The world population has grown significantly during this period, from around 5.3 billion people in 1990 to approximately 7.7 billion people in 2019. A larger population may naturally lead to an increase in the number of deaths.
2. Health Advances: Despite the increasing number of deaths, the death rate per 100,000 people may decrease due to advances in healthcare such as improved disease treatments, vaccination programs, and strengthened healthcare systems.
3. Epidemics and Epidemiology: Epidemics such as HIV/AIDS, tuberculosis, and influenza have caused millions of deaths, especially in the 1990s and 2000s. Additionally, the spread of infectious diseases may have resulted in sudden spikes in deaths and contributed to the overall increase in mortality.
4. Improvements in Diagnostic Technology and Reporting: The development of medical technology may have enhanced the ability to diagnose and report causes of death, leading to the recording of more deaths in databases.
5. Changes in Diet and Lifestyle: Changes in diet, lifestyle, and living environments may increase the risk of non-communicable diseases such as cardiovascular diseases, diabetes, and cancer, as well as deaths from non-infectious causes.

However, the uneven increase may reflect fluctuations in the impact of these factors over time, as well as the implementation of health strategies and policies in different periods.

Analyzing the causes of death is an important step in understanding the global burden of disease and developing effective interventions to reduce mortality. By identifying the leading causes of death, policymakers and health professionals can focus resources on the health issues that have the greatest impact on people's lives.



*Table 4.3: Code for visualizing data about the deaths by cause  
by year globally in the script.*

 *Table 4.4: The chart shows the total deaths by cause   
worldwide from 1990 to 2019*

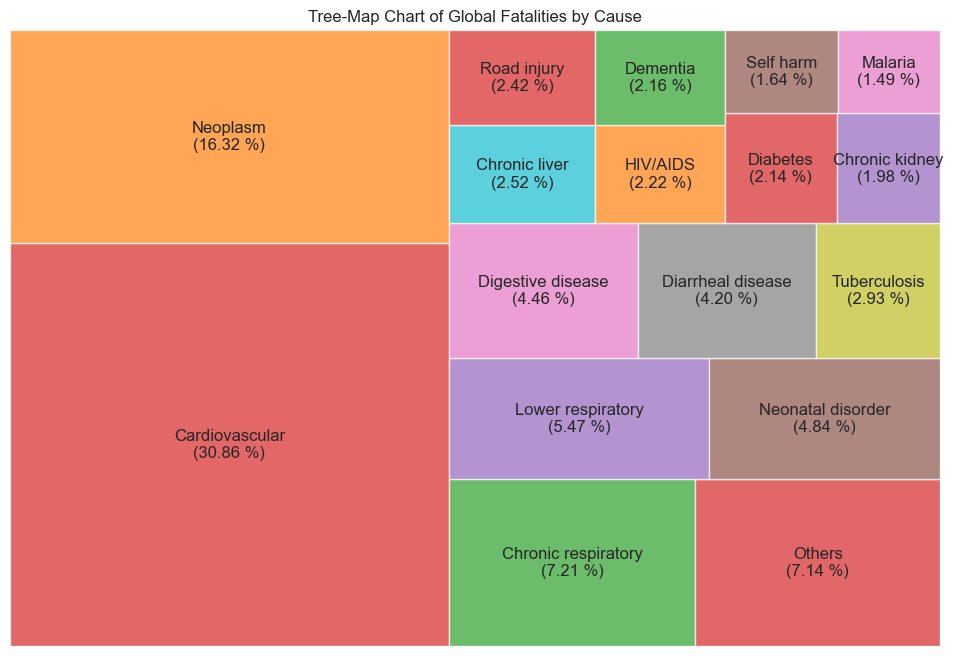
The chart provides an overview of the leading causes of death globally from 1990 to 2019. The data shows some notable trends, including an increase in mortality due to cardiovascular disease and stroke, declines in mortality from infectious diseases, and increases in mortality from cancer, Alzheimer's disease, and diabetes. These trends can be explained by a number of factors.

Including:

* Changing demographics.
* Lifestyle.
* Exposure to risk factors.

To have a more intuitive view of the rate of deaths by each cause, a tree map is a chart that may be the most reasonable choice.

*Table 4.5: Code for visualizing data about the deaths by cause  
by year globally with tree-map in the script.*

*Table 4.6: The chart shows the total deaths by cause   
worldwide from 1990 to 2019 with tree-map*

The Treemap chart depicts the number of deaths from leading causes globally from 1990 to 2019. The chart uses rectangles of different sizes to represent the death rate from each cause. The larger the size of the rectangle, the higher the mortality rate.

**Data analysis**

* Cardiovascular disease was the leading cause of death throughout this period, accounting for 30.86% of all deaths in 2019.
* Cancer is the second leading cause of death, accounting for 16.32% of all deaths in 2019.
* Lower respiratory disease is the third leading cause of death, accounting for 12.68% of all deaths in 2019.
* Digestive diseases are the fourth leading cause of death, accounting for 8.56% of all deaths in 2019.

**Discuss**

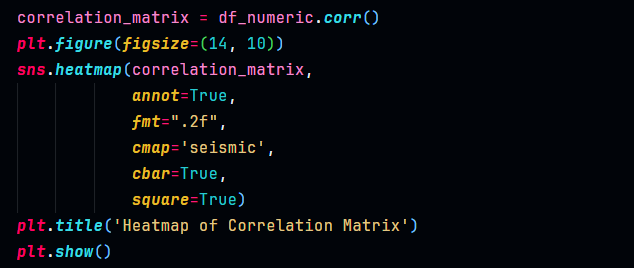
The increase in mortality from cardiovascular disease and cancer can be attributed to a number of factors, including aging populations, sedentary lifestyles, unhealthy diets and smoking. The decline in mortality from lower respiratory and digestive diseases can be attributed to improved sanitation, drinking water, and vaccination. The decline in infectious disease mortality can be attributed to the development of antibiotics, vaccines, and other public health measures.

**Compare with Chart 4.4**

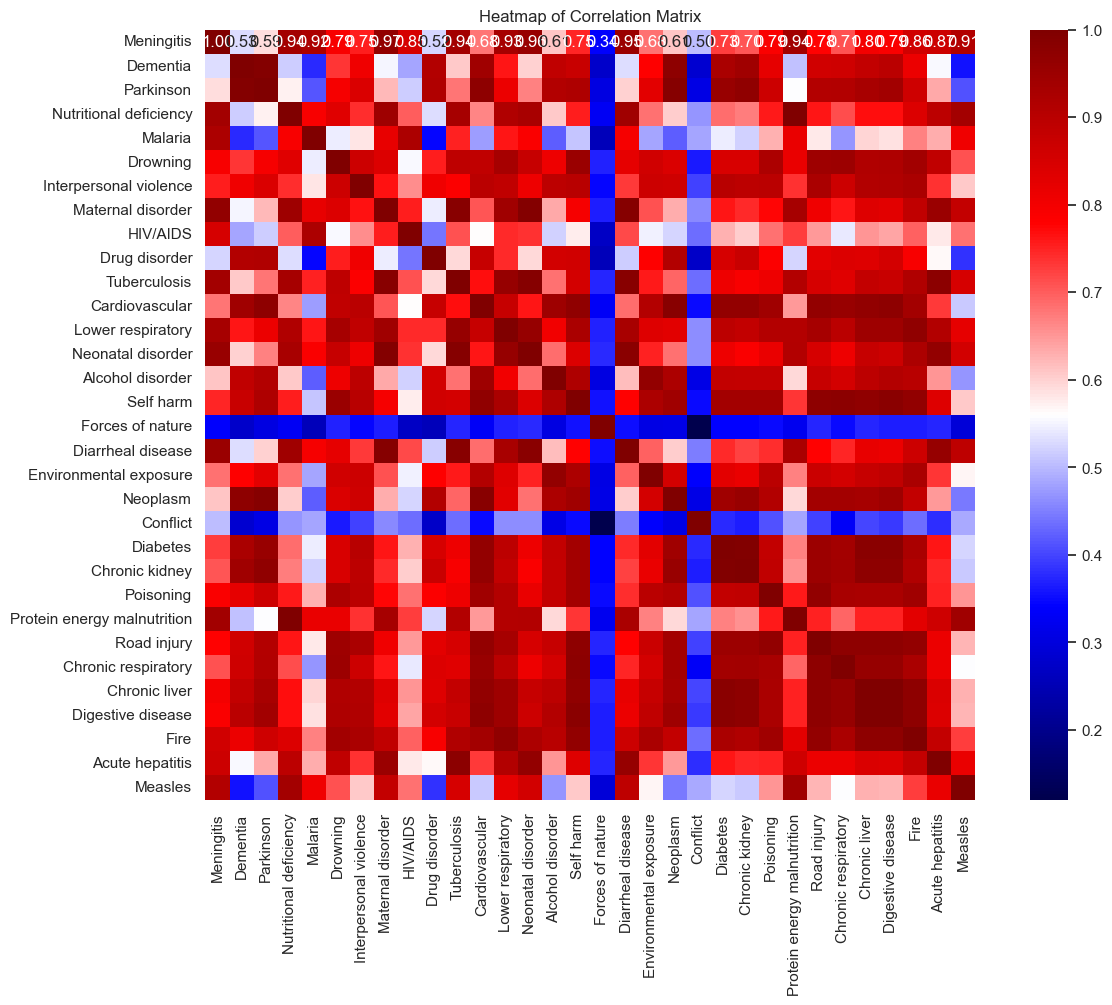
Treemap charts provide a visual way to compare death rates from different causes. Chart 4.4, on the other hand, provides a way to track mortality trends over time. Both charts provide valuable information about the global burden of disease.

## IV.2. Find relationships between causes of death and visualize with heatmap

From the above information, we can see that the number of deaths due to each cause is uneven, but do these causes have any interaction with each other? To clarify that, I calculated the ***correlation coefficient*** *(1)* of each pair of causes with each other and visualized it on a***heatmap*** *(2)* chart with a correlation matrix.

1. *A correlation coefficient is a statistical measure quantifying the linear relationship between two variables. It is used to analyze data sets or components of a random variable. Correlation coefficients range from -1 to +1, where values closer to ±1 indicate a stronger correlation, while 0 indicates no correlation.*
2. *In the correlation matrix, the heat map is a 2-dimensional data representation using colors to show the degree of correlation between pairs of variables. Each cell in the matrix is colored to correspond to the degree of correlation between the two corresponding variables. High correlation values are often represented by contrasting or light colors, while low correlation values are represented by dark colors. This helps identify correlation patterns in data in an intuitive and easy-to-understand way.*

*Table 4.3: Code for visualizing correlation matrix  
 by heatmap in the script.*

****

# Conclusion

# References