# Project documentation

Han Xie

03.11.2024

## Acknowledgment:

In this project, I have used different answers from stack overflow and the package documentation for ggplot, dplyr, tidyr, plotly, stringr. Those information helps me build the plots, and figure out the specific questions I have encountered.

I also acknowledge the help from Professor Priscilla Jiménez Pazmino and CSC 324 class mentor Elijah Mendoza in their help of finding problems and mentally modeling the shiny app.

I acknowledge the help from the Vivero Digital Fellows program in better designing the visualization and give advice about the graph's titles.

# Project Purpose

A straightforward way to assess the health status of a population is to focus on mortality – or concepts like child mortality or life expectancy, which are based on mortality estimates. A focus on mortality, however, does not take into account that the burden of diseases is not only that they kill people, but that they cause suffering to people who live with them. Assessing health outcomes by both mortality and morbidity (the prevalent diseases) provides a more encompassing view on health outcomes.

## Audience and Users

The audience of the project is for general audience. The shiny app with data visualization provides an accessible way for different audiences to select the information they want. The users could be scientists who would like to see how effective certain types of therapies work by looking at how the invention of certain medicine leads to the reduction of death numbers. The users could also be sociologists in viewing how different social conflicts cause the death of people. The users could also be general audiences who want to look at the death over years based on certain causes.

# Goal

As described in the purpose, the project aims to help people assess the health status of the population based on different causes of death. General data visualization helps users to view the trends of the variation in the death population.

# Data Description

There are 7 different datasets collected from 4 different sources.

## Source 1: data from WHO (2000_global.csv, 2010_global.csv, 2015_global.csv, 2019_global.csv)

The data from the World Health Organization (WHO) provides the cause-specific mortality with regard to gender and age information from 2000 - 2019. (data source: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death). It originally is one Excel file with different sheets (each sheet represents one year). Since r is not able to load the data in a roughly accurate format. I manually changed each of them into 4 separate csv files and cleaned data based on those 4 files.

Format: Each of the data frames involves 202 observations of 20 variables. All the observations are character variables. The variables were then converted to numbers information in r if they are numbers. The columns separate the total gender information and age information that separate by gender.

1. causes: describe different causes of diseases, character
2. total: total number of deaths, characters (convert to numeric data in the shiny app)
3. male: total male number of deaths, character (convert to numeric data in the shiny app)
4. female: total female number of deaths, character (convert to numeric data in the shiny app)
5. X0.28.days: total male number of deaths that in the age range of 0-28 days, character (convert to numeric data in the shiny app)
6. X1.59.months: total male number of death in the age range of 1-59 months, character (convert to numeric data in the shiny app)

7. X5.14.years: total male number of death in the age range of 5-14 years, character (convert to numeric data in the shiny app)
8. X15.29.years: total male number of death in the age range of 15-29 years, character (convert to numeric data in the shiny app)
9. X30.49.years: total male number of death in the age range of 30-49 years, character (convert to numeric data in the shiny app)
10. X50.59.years: total male number of deaths that in the age range of 50-59 years, character (convert to numeric data in the shiny app)
11. X60.69.years: total male number of deaths that in the age range of 60-69 years, character (convert to numeric data in the shiny app)
12. X70..years: total male number of death in the age range of 70 + years, character (convert to numeric data in the shiny app)
13. X0.28.days.1: total female number of deaths that in the age range of 0-28 days, character (convert to numeric data in the shiny app)
14. X1.59.months.1: total female number of deaths that in the age range of 1-59 months, character (convert to numeric data in the shiny app)
15. X5.14.years.1: total female number of death in the age range of 5-14 years, character (convert to numeric data in the shiny app)
16. X15.29.years.1: total female number of death in the age range of 15-29 years, character (convert to numeric data in the shiny app)
17. X30.49.years.1: total female number of death in the age range of 30-49 years, character (convert to numeric data in the shiny app)
18. X50.59.years.1: total female number of death in the age range of 50-59 years, character (convert to numeric data in the shiny app)
19. X60.69.years.1: total female number of death in the age range of 60-69 years, character (convert to numeric data in the shiny app)
20. X70..years.1: total female number of death in the age range of 70 + years, character (convert to numeric data in the shiny app)

## Source 2: Kaggle-Cause of Deaths around the World (Historical Data)

The data frame (csv file) from Cause of Deaths around the World (Historical Data) on the Kaggle website. (Data Source: https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world) It involves the data of different causes of death from 1990 to 2019 in different countries or regions.

Format: The data frame involves 34 columns and 6120 observations. The observations involve both numerics and character variables.

1. Country/Territory - Name of the regions, character
2. Code - region Codes, character
3. Year - Year of the Incident, numerics
4. Meningitis - No. of People died from Meningitis, numerics
5. Alzheimer's Disease and Other Dementias - No. of People died from Alzheimer's Disease and Other Dementias, numerics
6. Parkinson's Disease - No. of People died from Parkinson's Disease, numerics
7. Nutritional Deficiencies - No. of People died from Nutritional Deficiencies, numerics
8. Malaria - No. of People died from Malaria, numerics
9. Drowning - No. of People died from Drowning, numerics
10. Interpersonal Violence - No. of People died from Interpersonal Violence, numerics
11. Maternal Disorders - No. of People died from Maternal Disorders, numerics
12. HIV AIDS- No. of people died from HIV or AIDS, numerics
13. Drug Use Disorders - No. of People died from Drug Use Disorders, numerics
14. Tuberculosis - No. of People died from Tuberculosis, numerics
15. Cardiovascular Diseases - No. of People died from Cardiovascular Diseases, numerics
16. Lower Respiratory Infections - No. of People died from Lower Respiratory Infections, numerics
17. Neonatal Disorders - No. of People died from Neonatal Disorders, numerics
18. Alcohol Use Disorders - No. of People died from Alcohol Use Disorders, numerics
19. Self-harm - No. of People died from Self-harm, numerics
20. Exposure to Forces of Nature - No. of People died from Exposure to Forces of Nature, numerics
21. Diarrheal Diseases - No. of People died from Diarrheal Diseases, numerics
22. Environmental Heat and Cold Exposure - No. of People died from Environmental Heat and Cold Exposure, numerics
23. Neoplasms - No. of People died from Neoplasms, numerics
24. Conflict and Terrorism - No. of People died from Conflict and Terrorism, numerics
25. Diabetes Mellitus - No. of People died from Diabetes Mellitus, numerics

26. Chronic Kidney Disease - No. of People died from Chronic Kidney Disease, numerics
27. Poisonings - No. of People died from Poisoning, numerics
28. Protein-Energy Malnutrition - No. of People died from Protein-Energy Malnutrition, numerics
29. Road Injuries – No. of people died from road injuries, numerics
30. Chronic Respiratory Diseases - No. of People died from Chronic Respiratory Diseases, numerics
31. Cirrhosis and Other Chronic Liver Diseases - No. of People died from Cirrhosis and Other Chronic Liver Diseases, numerics
32. Digestive Diseases - No. of People died from Digestive Diseases, numerics
33. Fire, Heat, and Hot Substances - No. of People died from Fire or Heat or any Hot Substances, numerics
34. Acute Hepatitis - No. of People died from Acute Hepatitis, numerics

## Source 3: Kaggle- World Population Analysis

The data frame (csv file) about the world population analysis is sourced from Kaggle. (data source: https://www.kaggle.com/code/hasibalmuzdadid/world-population-analysis/notebook) Data involves the population of the world based on different regions and continents.

Format: There are 234 observations of 17 variables. The observations involve both numerics and character variables.

1. Rank: Rank by population, numerics, character
2. CCA3: 3 digit Country/Territories code, character
3. Country: Name of the Country/Territories, character
4. Capital: Name of the Capital, character
5. Continent: Name of the Continent, character
6. 2022 Population: Population of the Country/Territories in the year 2022, numerics
7. 2020 Population: Population of the Country/Territories in the year 2020, numerics
8. 2015 Population: Population of the Country/Territories in the year 2015, numerics
9. 2010 Population: Population of the Country/Territories in the year 2010, numerics
10. 2000 Population: Population of the Country/Territories in the year 2000, numerics
11. 1990 Population: Population of the Country/Territories in the year 1990, numerics
12. 1980 Population: Population of the Country/Territories in the year 1980, numerics

13. 1970 Population: Population of the Country/Territories in the year 1970, numerics
14. Area ($km^2$): Area size of the Country/Territories in square kilometers, numerics
15. Density (per $km^2$): Population density per square kilometer, numerics
16. Growth Rate: Population growth rate by Country/Territories, numerics
17. World Population Percentage: The population percentage by each Country/Territories, numerics

## Source 4: Maps package: world map

I have used the maps package in the r library to extract the longitude and latitude for the world map. (Data source: https://cran.r-project.org/web/packages/maps/index.html) Data involves the geographical location of each region.

Format: There are 98977 observations of 6 variables, The observations involve both numerics and character variables.

1. long: longitude information, numerics
2. lat: latitude information, numerics
3. group: label for countries, numerics
4. order: the order of each observation, numerics
5. region: the region name, characters
6. subregion: the subregion name, characters

# Questions

I am trying to understand the death information for each of the causes. Here, I would use Malaria as an example to demonstrate the specific questions, but it could apply to different causes listed.

1. What are the gender and age ratios for the death of Malaria?
2. Which country has the greatest number of deaths?
3. What is the trend of death numbers over time?
4. What is the mortality rate due to Malaria?

# Insights

By looking at the data visualization I did, I can find that:

1. The total death number due to Malaria is 21768985 from 1990 to 2019.
2. While looking at each continent, Africa has the greatest number of people who died of Malaria (more than 18 million); The second greatest is Asia where more than 3 million people died due to malaria.
3. The ratio for males and females is roughly equal (49.9 to 50.1)
4. Malaria affects children the most, there are more than 75% of deaths are between 1 to 59 months (0 to 4 years old), and 6.54% of deaths are between 5 to 14 years old.
5. The country that has been affected the most is Nigeria, with a total amount of 6.4 million. The second affected country is India, with an amount of 2.44 million.
6. The total mortality trend due to Malaria increased from 1990 to 2003 and then decreased from 2003 to 2019. This might be due to the development of medicine which functionally prevents the increment of mortality.

# the process to make work reproducible

1. I have used rename() to keep the column name to be the same name in population data and the cause of death data.

```
population1 <- read.csv("world_population.csv") %>% rename("Code"="CCA3")
```

2. I have used string replace to ensure the same name in the map and the cause of death

   ```
   data. data1$Country.Territory <- str_replace(data1$Country.Territory, "United States", "USA")
   ```

3. I have merged the world population and cause of death data by code and country/region name

   ```
   data1 <- merge(death, population, by = c("Code", "Country.Territory"))
   ```

4. I have used mutate to make the variables numerics

   ```
   summary <- summary %>% mutate(totalNum = as.numeric(gsub(",","", summary$total))) %>% mutate(maleNum = as.numeric(gsub(",","", summary$male))) %>% mutate(femaleNum = as.numeric(gsub(",","", summary$female))) %>% mutate(M_0_28_days = as.numeric(gsub(",","", summary$X0.28.days))) %>% mutate(F_0_28_days = as.numeric(gsub(",","", summary$X0.28.days.1))) %>% mutate(M_1_59_mons = as.numeric(gsub(",","",
   ```

summary$X1.59.months))) %>% mutate(F_1_59_mons = as.numeric(gsub(",","", summary$X1.59.months.1))) %>% mutate(M_5_14_years = as.numeric(gsub(",","", summary$X5.14.years))) %>% mutate(F_5_14_years = as.numeric(gsub(",","", summary$X5.14.years.1))) %>% mutate(M_15_29_years = as.numeric(gsub(",","", summary$X15.29.years))) %>% mutate(F_15_29_years = as.numeric(gsub(",","", summary$X15.29.years.1))) %>% mutate(M_30_49_years = as.numeric(gsub(",","", summary$X30.49.years))) %>% mutate(F_30_49_years = as.numeric(gsub(",","", summary$X30.49.years.1))) %>% mutate(M_50_59_years = as.numeric(gsub(",","", summary$X50.59.years))) %>% mutate(F_50_59_years = as.numeric(gsub(",","", summary$X50.59.years.1))) %>% mutate(M_60_69_years = as.numeric(gsub(",","", summary$X60.69.years))) %>% mutate(F_60_69_years = as.numeric(gsub(",","", summary$X60.69.years.1))) %>% mutate(M_70_years = as.numeric(gsub(",","", summary$X70..years))) %>% mutate(F_70_years = as.numeric(gsub(",","", summary$X70..years.1)))

5. I have used mutate to ensure the causes name to be the same

   summary <- summary %>% mutate(causes = case_when( causes == "Protein-energy malnutrition" ~ "Protein.Energy.Malnutrition", causes == "Iodine deficiency" ~ "Nutritional.Deficiencies", causes == "Vitamin A deficiency" ~ "Nutritional.Deficiencies", causes == "Iron-deficiency anaemia" ~ "Nutritional.Deficiencies", causes == "Other nutritional deficiencies" ~ "Nutritional.Deficiencies", causes == "Alzheimer disease and other dementias" ~ "Alzheimer.s.Disease.and.Other.Dementias", causes == "Parkinson disease" ~ "Parkinson.s.Disease", causes == "Interpersonal violence" ~ "Interpersonal.Violence", causes == "Maternal conditions" ~ "Maternal.Disorders", causes == "HIV/AIDS" ~ "HIV.AIDS", causes == "Drug use disorders" ~ "Drug.Use.Disorders", causes == "Rheumatic heart disease" ~ "Cardiovascular.Diseases", causes == "Hypertensive heart disease" ~ "Cardiovascular.Diseases", causes == "Ischaemic heart disease" ~ "Cardiovascular.Diseases", causes == "Stroke" ~ "Cardiovascular.Diseases", causes == "Cardiomyopathy, myocarditis, endocarditis" ~ "Cardiovascular.Diseases", causes == "Other circulatory diseases" ~ "Cardiovascular.Diseases", causes == "Drug use disorders" ~ "Drug.Use.Disorders", causes == "Lower respiratory infections" ~ "Lower.Respiratory.Infections", causes == "Preterm birth complications" ~ "Neonatal.Disorders", causes == "Birth asphyxia and birth trauma" ~ "Neonatal.Disorders", causes == "Neonatal sepsis and infections" ~ "Neonatal.Disorders", causes == "Other neonatal conditions" ~ "Neonatal.Disorders", causes == "Alcohol use disorders" ~ "Alcohol.Use.Disorders", causes == "Self-harm" ~ "Self.harm", causes == "Natural disasters" ~ "Exposure.to.Forces.of.Nature", causes == "Diarrhoeal diseases" ~

"Diarrheal.Diseases", causes == "Malignant neoplasms" ~ "Neoplasms", causes == "Other neoplasms" ~ "Neoplasms", causes == "Interpersonal violence" ~ "Conflict.and.Terrorism", causes == "Collective violence and legal intervention" ~ "Conflict.and.Terrorism", causes == "Diabetes mellitus" ~ "Diabetes.Mellitus", causes == "Alcohol use disorders" ~ "Chronic.Kidney.Disease", causes == "Chronic kidney disease due to diabetes" ~ "Chronic.Kidney.Disease", causes == "Road injury" ~ "Road.Injuries", causes == "Respiratory diseases" ~ "Chronic.Respiratory.Diseases", causes == "Cirrhosis of the liver" ~ "Cirrhosis.and.Other.Chronic.Liver.Diseases", causes == "Other neoplasms" ~ "Neoplasms", causes == "Digestive diseases" ~ "Digestive.Diseases", causes == "Fire, heat and hot substances" ~ "Fire..Heat..and.Hot.Substances", causes == "Hepatitis" ~ "Acute.Hepatitis", TRUE ~ causes))

6. I have used group_by and summarise to find the total population over years

   group_by(causes)%>% summarize_all(.funs = sum)

7. calculate the total population of death from age, gender facet

   mutate(T_0_28_days = M_0_28_days + F_0_28_days) %>% mutate(T_1_59_mons = M_1_59_mons+ F_1_59_mons) %>% mutate(T_5_14_years = M_5_14_years + F_5_14_years) %>% mutate(T_15_29_years = M_15_29_years + F_15_29_years) %>% mutate(T_30_49_years = M_30_49_years + F_30_49_years) %>% mutate(T_50_59_years = M_50_59_years + F_50_59_years) %>% mutate(T_60_69_years = M_60_69_years + F_60_69_years) %>% mutate(T_70_years = M_70_years + F_70_years)

8. For each of the plots, I have used different filters, group_by, select, mutate, summarise, pivot_longer, etc. functions to be able to get the expected data inputs.

e.g. 1. data for the bar chart:

filtered_data <- reactive({ subset(data1, Year >= input$year_range[1] & Year <= input$year_range[2]) })

continent_death <- reactive({ filtered_data() %>% select(-c("Code","Country.Territory", "Year", "Capital"))%>% group_by(Continent)%>% summarise_all(.funs=sum)%>% pivot_longer(!Continent, names_to = "causes", values_to="death_number")%>% mutate(death_number_per_100_000_0 = death_number/1000000) })

e.g. 2. data for the pie chart

cause <- sumorder()$causes[event_data("plotly_click")$y] findCause <- age_sex_info%>% select(c(maleNum, femaleNum, causes))%>% filter(causes == cause)%>% pivot_longer(!causes, names_to = "right_num", values_to = "num")

e.g. 3. data for the map:

mapCause <- toString(sumorder()$causes[event_data("plotly_click")$y]) country_cause <- country_cause()%>% filter(causes == mapCause)

e.g. 4. data for the scatterplot: mapCause <- toString(sumorder()$causes[event_data("plotly_click")$y])

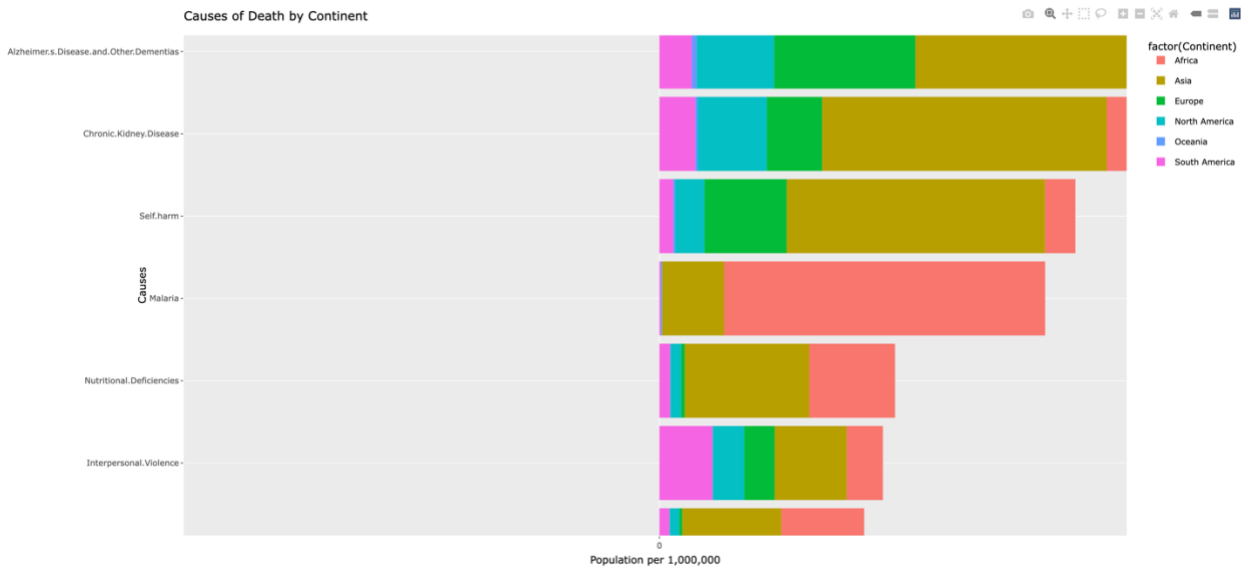
```
country_code <- as.numeric(event_data("plotly_hover")[2])+1
The_code <- countryName()$Code[country_code]


filteredCause <- country_cause()%>%
  filter(Code == The_code)%>%
  filter(causes == mapCause)%>%
  select(-c("death_number"))


overYear <- filtered_data()%>%
  filter(Code == The_code)%>%
  select(c("Year", mapCause, "Code"))%>%
  left_join(filteredCause, by = "Code")%>%
  rename("death_number" = mapCause)%>%
  select(c("Year","death_number"))
```

# design decisions

## bar chart

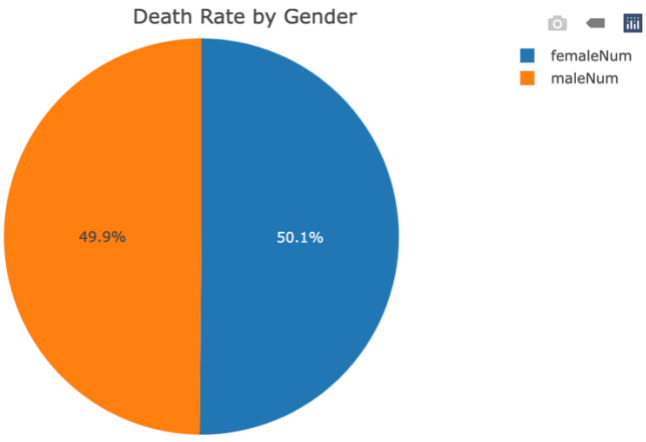



What? The bar chart shows different causes over population with different continents information

Why? The chart could help visualize the death number, especially for different continents.

How? While looking at each continent, Africa has the greatest number of people who died of Malaria (more than 18 million); The second greatest is Asia where more than 3 million people died due to malaria.

# pie chart 1

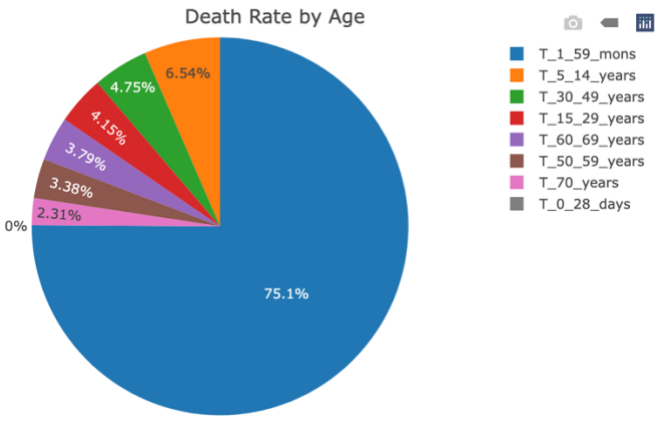



What? The pie chart shows the gender ratio

Why? The pie chart allows me to find the ratio directly

How? The ratio for males and females is roughly equal (49.9 to 50.1)

# pie chart 2

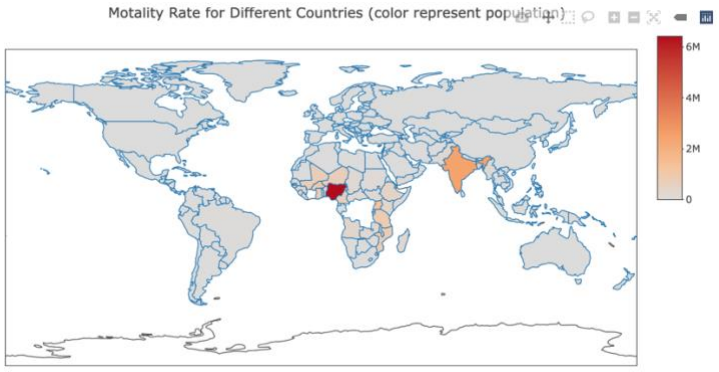



What? The pie chart shows the age ratio

Why? I want to see the age distribution of death

How? Malaria affects children the most, there are more than 75% of deaths are between 1 to 59 months (0 to 4 years old), and 6.54% of deaths are between 5 to 14 years old.

# Map

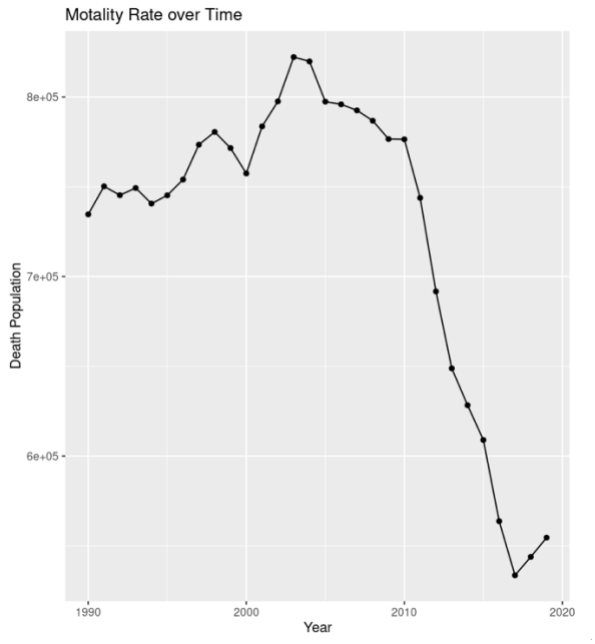



What? The map shows how the mortality rate over countries

Why? I want to find the countries that have been affected the most.

How? From the graph, the country that has been affected the most is Nigeria, with a total amount of 6.4 million. The second affected country is India, with an amount of 2.44 million.

# Scatterplot




What? The scatterplot shows the population over years.

Why? I want to see the trend of variation of mortality over time

How? The total mortality trend due to Malaria increased from 1990 to 2003 and then decreased from 2003 to 2019. This might be due to the development of medicine which functionally prevents the increment of mortality.

# Limitation with potential improvement

1. Lack of information about the death cause of exposure of heat and cold for the two pie chats since I cannot find related information about that.
2. The gender is only limited to the two types of genders which is not inclusive enough for LGBTQ group
3. Since I am only able to find one way to use one click, if user accidentally click on the other visualizations, the graph will show as error. I do not have a good way to solve it for now

4. Not able to unselect the causes for the bar chart

# Reference


1. CSC 324 class page: https://jimenezp.cs.grinnell.edu/Courses/CSC324/2024Sp/schedule/
2. Interactive plot: https://shiny.posit.co/r/articles/build/plot-interaction/
3. WHO data: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death
4. Kaggle data:
    1. world population analysis: https://www.kaggle.com/code/hasibalmuzdadid/world-population-analysis/notebook)
    2. Cause of Deaths around the World (Historical Data): https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world)
5. World map data: https://cran.r-project.org/web/packages/maps/index.html


# Appendix

1. A particularly nice piece of code, for example, animations, extensions.
   I have involved the interactive plots and the interlock between the plots in my code. That allows users to click on one of the plots to link to filter information for another graphs. By hovering on the plot to see the changes in another plots. In order to reach the interaction, I have used many different packages includes plotly, htmlwidgets, tidyverse, etc.

2. Combine various datasets.
   I have combine 7 different datasets from 4 different sources as described in the data description above. And I need to do lots of data manipulation in order to makes them works for what types of data visualization I want to make.

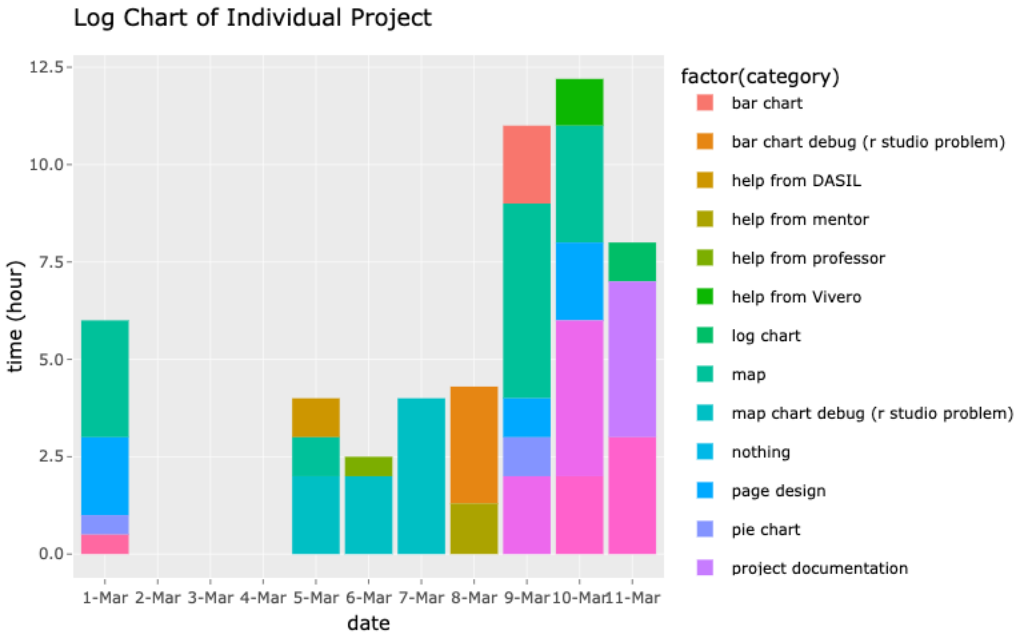
3. An outstanding recording (with good audio) within the time range.
   The recording is within the time range.

4. Feedback report of your visit to the Vivero Digital Fellows program.
   I went to the Vivero Digital Fellows program, and here is a feedback report:
    1. They give me advice on how to make the variables description more precise. For example, they advise me to change the name from “malenum” to “male” in the first pie chart so that users could better understand what they are looking for.

Besides, they also advise to change for the ages, however, since I am not able to change it directly (I cannot find a way to put the number in the very front for a variable name)

2. They help me add description on user instruction. For example, let user select the bar to select the causes. And remind user to hover on a region to see the data on that region.
3. They helps me to add some nice titles for each of the visualization.
4. They also let me think if I can make the histogram larger so that the user would be easier to select the causes and it would look less overwhelming when first open the webpage. I think those are really helpful advice!

5. Well-designed time log and chart. (Facilitate identification of activities, tracking productivity, time spent in different activities, visualization of project timeline)




The log chart above shows the time I spent in different activities with different colors, and the x axis is the timeline. The graph in the rmd file allows to hover on and see the exact time spent on each day with each tasks.