Assignment Coversheet

**IOE – FACULTY OF EDUCATION AND SOCIETY**

|  |  |
| --- | --- |
| Candidate Number:   4 letters plus 1 digit, e.g. ABCD1. Do not use your name. | PPZM2 |
| Programme Title: Programme you are enrolled on. | Computational Social Science |
| Module Title: | SOCS0100 |
| Assignment Title: | First Assignment |
| Word Count: See module guide for exclusions. | 1351 |
| Generative Artificial Intelligence (GenAI) use:  Have you made use of AI at any stage of your assignment preparation?  Yes  No  **If yes**, please list each tool you used and include a statement of how you used it - for further information, please refer to: [Engaging with AI in your education and assessment | Students - UCL – University College London](https://www.ucl.ac.uk/students/exams-and-assessments/assessment-success-guide/engaging-ai-your-education-and-assessment) | |

Student Declaration:

By submitting this assignment, I confirm that all the work is my own unless collaboration has been specifically authorised. I understand that any form of [Academic Misconduct](https://www.ucl.ac.uk/students/exams-and-assessments/academic-integrity) is strictly prohibited, including the use of essay mills, homework help sites, plagiarism, collusion, falsification, impersonation, undeclared GenAI use or any other action which might give me an unfair advantage.

​ ← **Tick this box to confirm**

Use for teaching:

My assignment may be used in anonymised form as an exemplar so that future students can use it as a learning tool.

​ ​ ← **Opt Out: Do NOT use my anonymised submission in teaching.**

Only complete if you submit after the original deadline:

I have completed the Delayed Assessment Scheme task in Portico to use one of my three Delayed Assessment Permits.

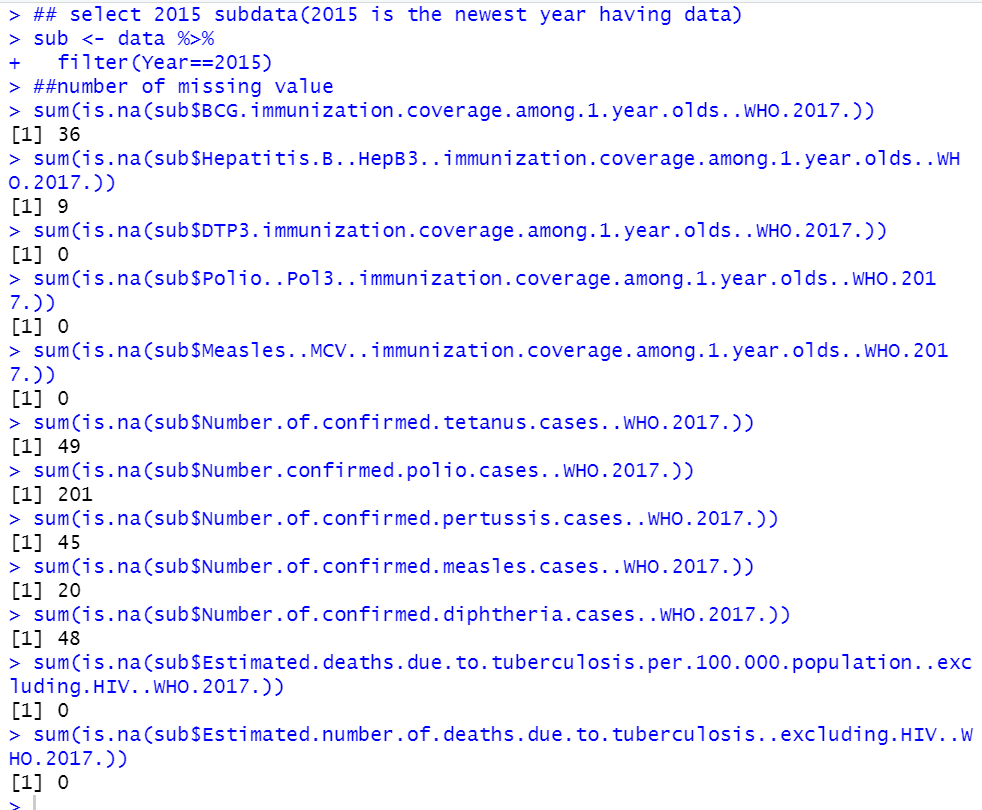
I have applied for EC and received a decision; my new deadline is: Click or tap to enter a date.

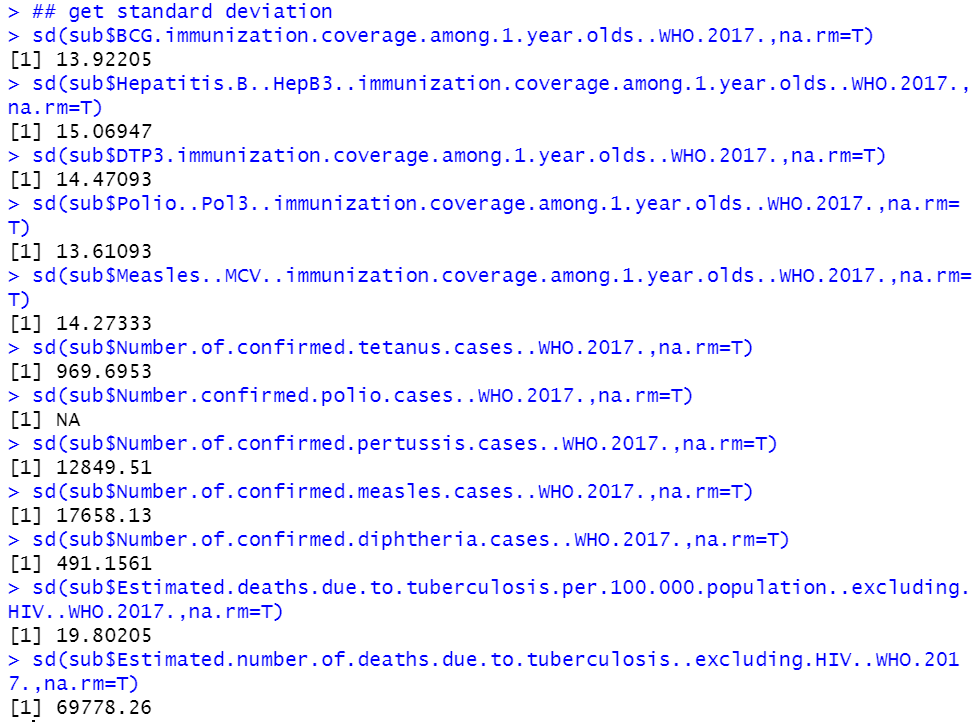
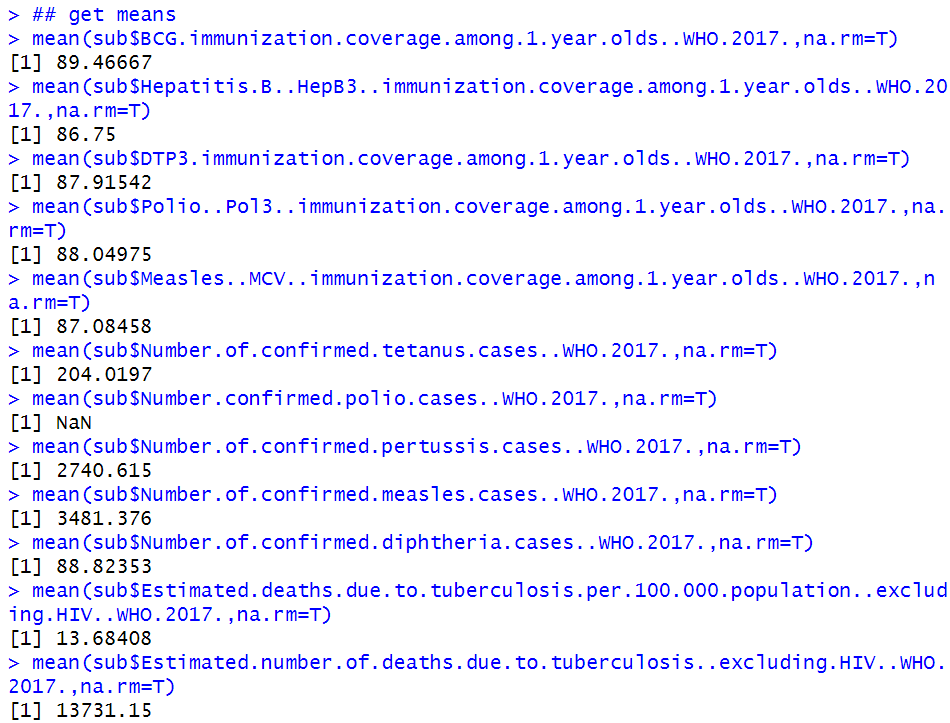
I have applied for EC, but not yet received a decision from the panel yet.

The dataset provided by the World Health Organization (WHO) in 2017 contains statistics on vaccination coverage and disease burden, offering valuable insights into global vaccination rates and their impact on disease prevalence. This dataset holds significant historical importance for public health research and policy-making, as it can be used to assess vaccination coverage and its correlation with the burden of preventable diseases.

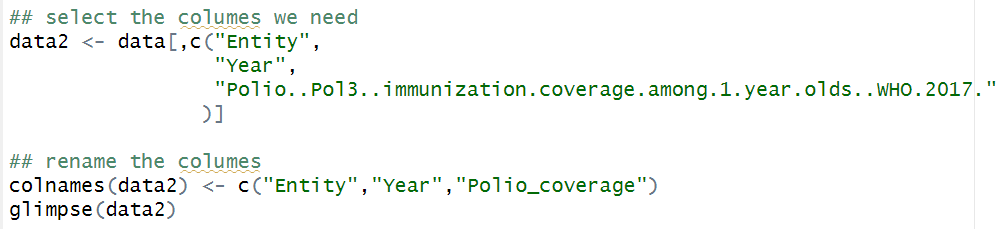
The dataset includes variables such as entity, year, vaccination coverage for five vaccine types, occurrences of five diseases, estimated tuberculosis deaths per 100,000 people (excluding HIV), and estimated deaths from tuberculosis (excluding HIV).

The filter() function was used to select data for 1980, stored in sub. It was found that there are 36 missing values in the BCG immunization coverage among 201 entities. The mean and standard deviation of each vaccination rate and disease variable were calculated, revealing that the mean BCG immunization coverage among 1-year-olds is 89.47, with a standard deviation of 13.92. The code calculates the missing values, mean, and standard deviation using the mentioned functions.

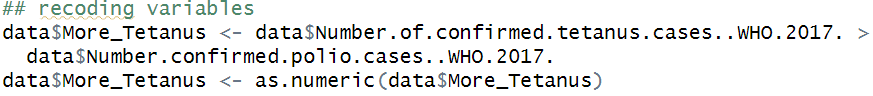




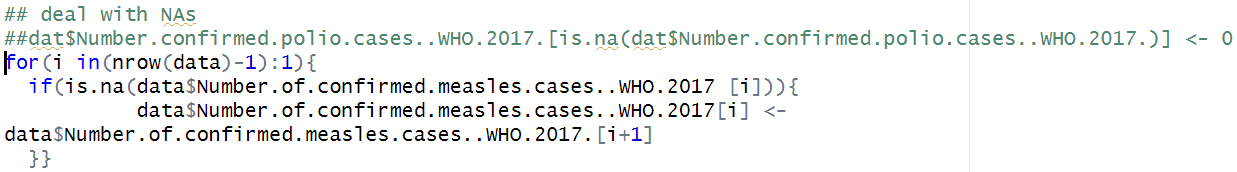
· I delete variables that I do not want to use, and renaming picked variables which names are too long and complex so that I can have tidy data frame and focused on less variables by selecting variables. I kept only Poli.. Pol3.. immunization. coverage. among.1.year.olds.. WHO. 2017. And renamed it as Polio\_Pol3.



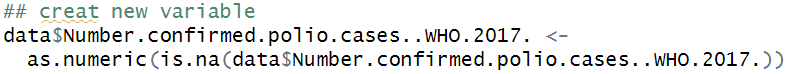
· I set the More\_Tetanus variable based on the comparison of the number of tetanus and polio cases. tetanus.cases >polio.cases Returns a Boolean vector indicating whether the number of tetanus cases in each row is greater than the number of polio cases. TRUE means more cases of tetanus and FALSE means more or equal cases of polio. The as.numeric( ) function converts Boolean values to numeric values, where TRUE is converted to 1 and FALSE is converted to 0. After this is done, the More\_Tetanus column will contain either 1 (more tetanus cases) or 0 (more or equal polio cases). This new variable will help health authorities prioritize which diseases to prevent and control, and develop relevant immunization programs.



· Generally, replacing NA to 0 is acceptable solution. But here, no information does not mean there is no one get disease. I decide to fill the cell upwards-replacing the missing value with the value of the next row. (is.na (measles. cases [i] ) ) function checks if the row i of column Number of confirmed measles cases WHO 2017 is a missing value (NA). If the value is missing, the condition is TRUE, and the statement in the code block— {measles. cases [i] <- measles. cases [i+1] } will be executed which is replacing the missing value with the value of the next row. This process starts from the second to last row. The value of i (the loop control variable or number of loops) is reduced from nrow(data) -1 (the second-to-last row) to 1 (the first row), implementing bottom-up traversal.

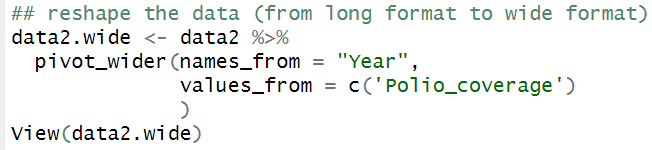


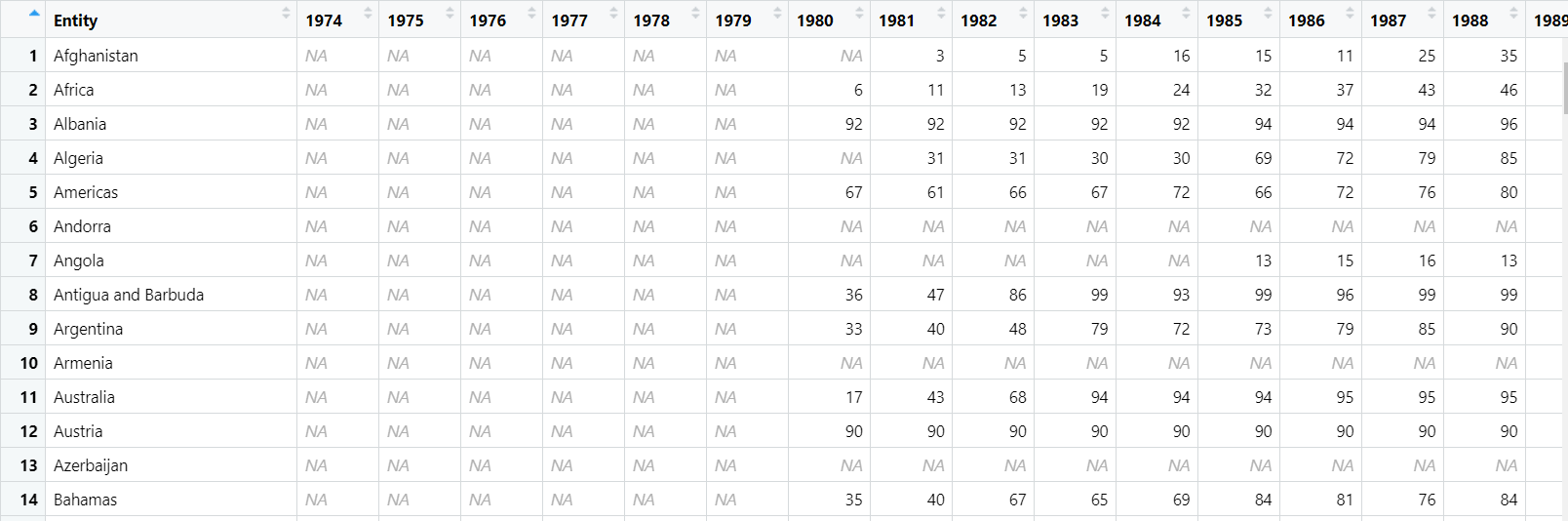
· This line of code will replace the value of column polio. cases with 0 or 1, where 1 means NA and 0 means there is data. This is actually creating a new variable to flag the presence of missing data, rather than filling in the NA value. I only interested in does variable have missing value, and I want to do something which can make further knowing how many NA the variable has quickly, instead of the accurate case number, so I make this new variable.



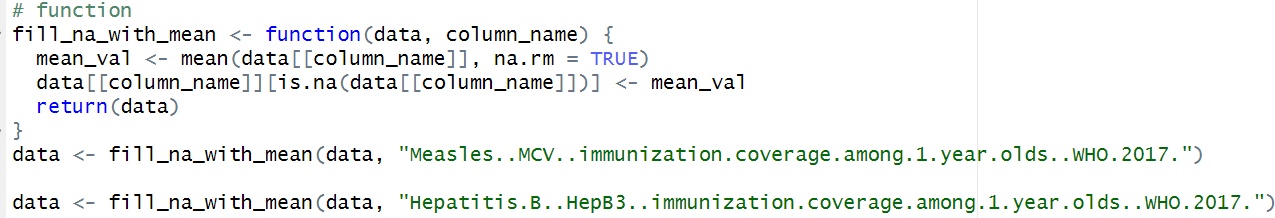
· I used the pivot\_wider function to convert the data frame from long to wide format, merging multiple rows into one and transforming some column values into new columns. The names\_from parameter specifies the source for the new column names, while the values\_from parameter indicates which column values fill these new columns (in this case, the Polio\_coverage column).

There are two reasons for this transformation. First, long format makes it difficult to compare cells since data for one variable is spread across multiple rows, leading to repeated information. In contrast, wide format presents different variable categories as separate columns, allowing for easier side-by-side comparison of values across time periods or conditions. Additionally, many visualization functions, such as scatter plots, require data in wide format for better convenience, especially with multiple classes and features.

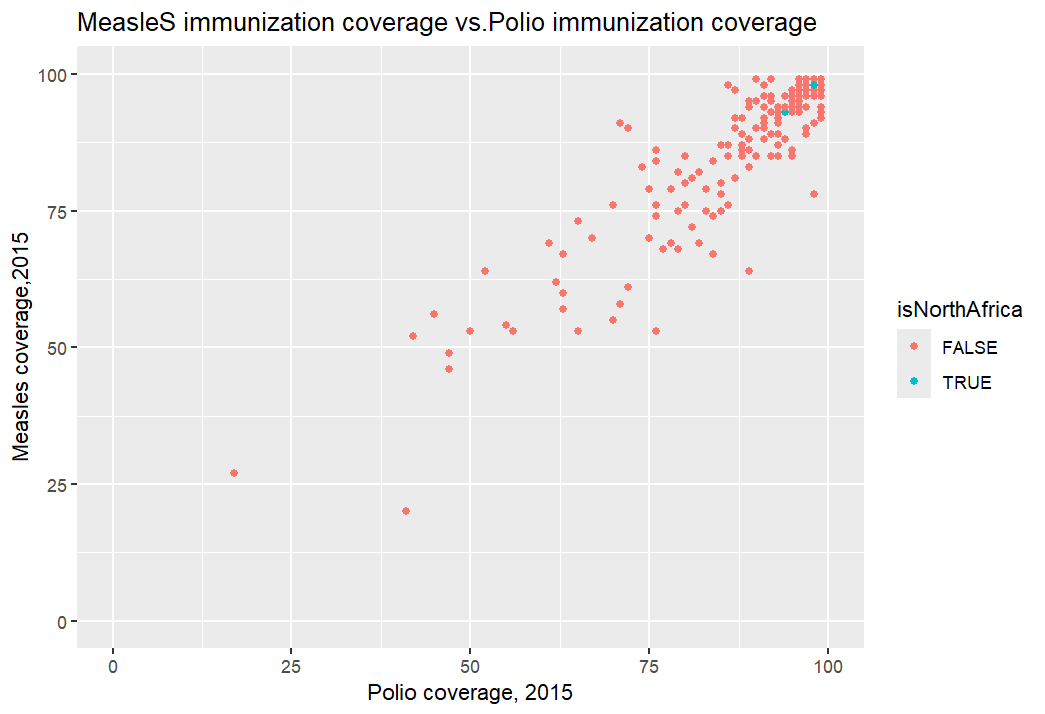




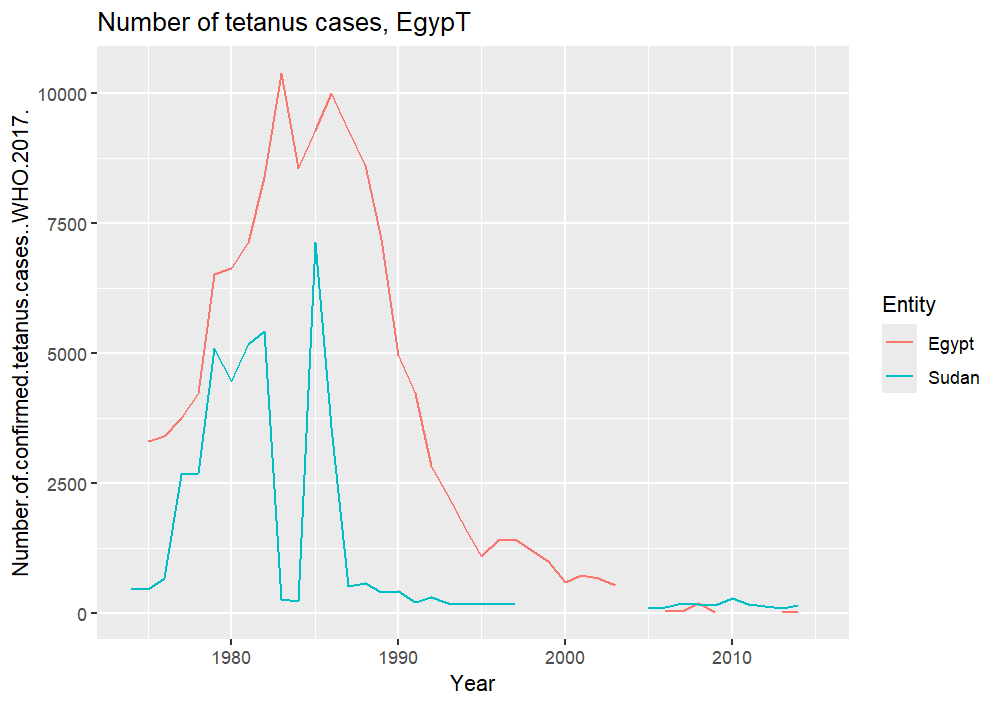
· I handle the missing values for Measles and Hepatitis coverage with the function. Missing values are a common problem in data analysis, and they can lead to inaccurate results. By filling in missing values, the integrity of the data set can be maintained and the accuracy of the analysis improved. And using this function can easily handle the missing value problem in different columns at the same time, and pass the column name as a parameter, which improves the efficiency of data processing

****

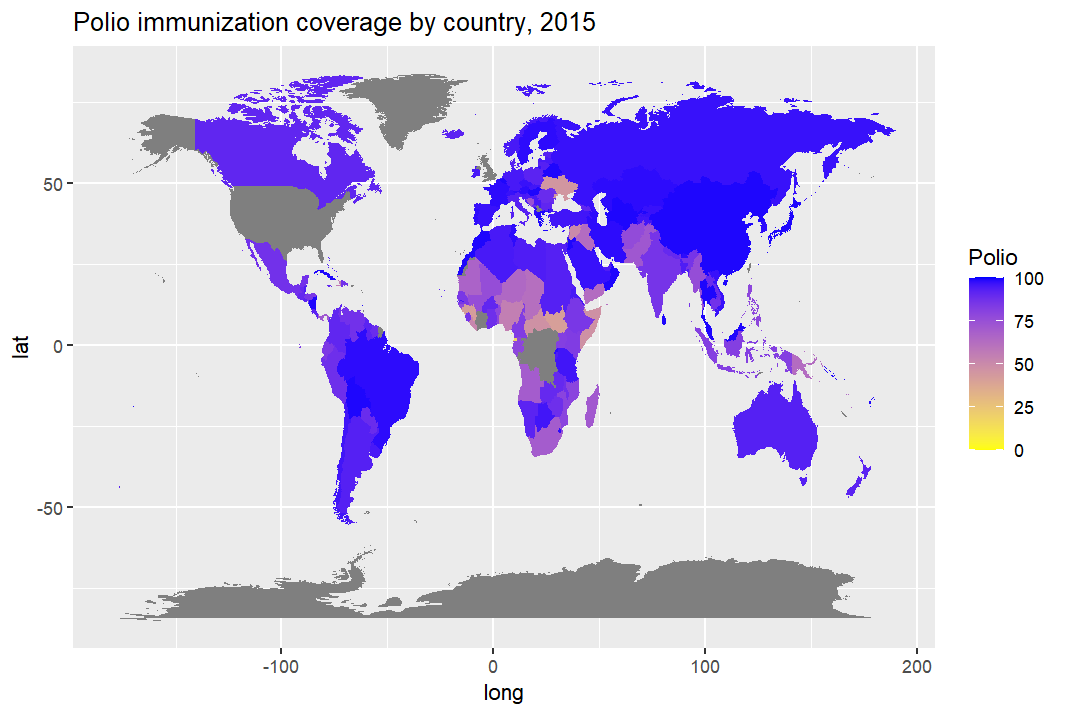
This scatter plot compares polio and measles immunization coverage among 1-year-olds in 2015. The X-axis shows polio coverage, while the Y-axis shows measles coverage. Blue dots represent North African countries, and red dots represent other countries. North African countries generally achieved high coverage in both vaccines, and most other countries also showed high coverage, with few displaying low rates for both vaccines. Vaccination rates for both vaccines are similar in most countries. By using color to distinguish North African countries from other countries, the chart directly compares immunization practices in different regions. Scatter plots are clear and easy to understand and can be quickly understood by non-specialists. The scatter plot here effectively shows a country's attitude towards two types of vaccination, and can be an entry point for an in-depth study of a country's health system. In visualization process, I specifically set x and y axes to range 0-100 to visually compare two vaccinations in a country.



I plotted the annual trend in the number of tetanus cases in Egypt and Sudan. Before 2000, the number of tetanus cases in the two countries fluctuated greatly, and the number of cases in Egypt was always more than Sudan, and even twice as many. After 1985, the quantity has a downward trend, and after 2005, it approaches 0, which becomes a fluctuation trend. The line chart can see all the data of a variable in multiple countries, so it is more convenient to compare the data of different countries in the same year and the change trend. There is support for exploring the impact of vaccine rollout and disease outbreaks on the number of cases. This can provide a basis for further research on the effectiveness of national policies or interventions. Among them, I use replaceNA function to take the average of the last two years of data to deal with the missing values. The use of functions makes it easy for me to switch between multiple countries without having to write code repeatedly. Function code is relatively simple, you can understand the function from the name, easy to understand the code logic.



Using ggplot2 and map\_data, a world map was created to display polio immunization coverage by country in 2015. Blue and yellow shades highlight areas with high and low coverage, respectively. Pale yellow regions, mainly in Africa, Southeast Europe, the Middle East, South, and West Asia, indicate lower coverage levels, while North America and Europe show generally high national coverage. This geographic view highlights regional immunization differences and also identifies countries with missing data, guiding further data exploration.



I think ChatGPT helping soliciting its code refinement in data wrangling and visualization is not acceptable. Because this task requires data sorting, wrangling and visualization. ChatGPT helps me check for errors in code that doesn't run, which is helpful because when I'm tired, I can't tell if I'm case-sensitive or if I'm missing parentheses and quotes，In the third visualization I couldn't figure out what to draw, but ChatGPT suggested map and provide the code frame. ChatGPT sometimes provided repetitive or difficult code, which confused me so much that I could only look up the code keywords that might be used in the materials in the class. ChatGPT quickly responded to my questions and summarized the knowledge points involved in the questions, which was helpful for me to find the materials in the class.