

STA Honours - Statistical Computing - Assignment 2

Department of Statistical Science, University of Cape Town

2024-02-29

Instructions

Complete this assignment in either Rmarkdown or Quarto.

Hand in both the Rmarkdown / Quarto file and the compiled pdf file.

Due Date: 5 March 2024

1 Gapminder

There are 3 gapminder data sets:

(2)

- `lex.csv` with life expectancy
- `gdp_pcap.csv` with GDP per Capita income
- `Data Geographies - v2 - by Gapminder.xlsx` with data on which region each country belongs to (four regions).

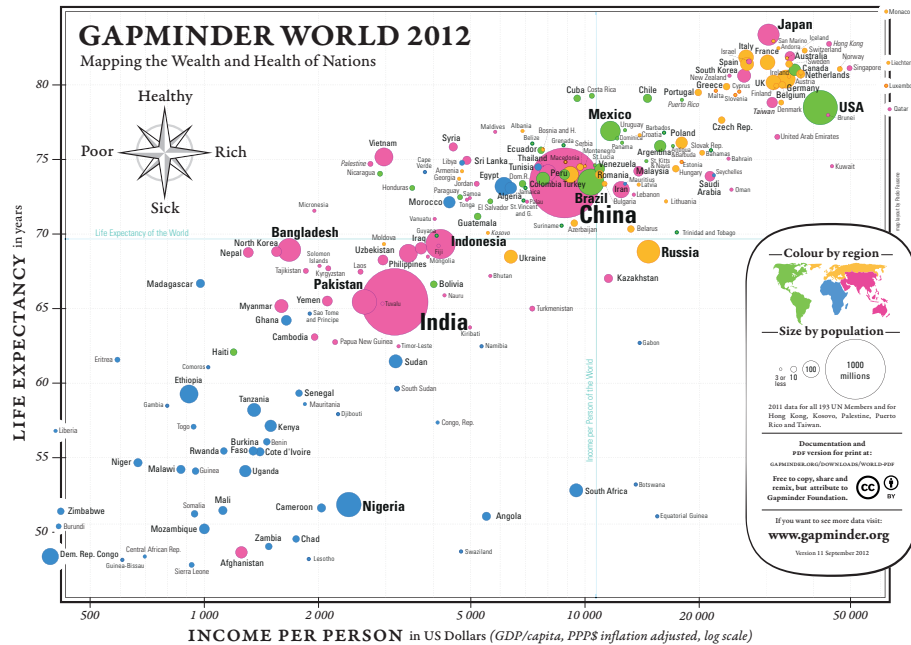


Figure 1: <https://www.gapminder.org/downloads/gapminder-world-2012-pdf/>

- Using `ggplot`, create a scatter plot of life expectancy against GDP per capita. Colour by region, size of points by population. Use 2019 data only. Add country names (see Figure 1). (6)
- Calculate average life expectancy per region, as well as number of countries per region. Summarize this in a table, sorted by decreasing life expectancy. For 2019. (3)
- Plot the average life expectancy over time, per region (1950 - 1919). (3)
- Compare the distribution of life expectancy in 2019 for the different regions. Choose a suitable type of plot. (4)
- How many different countries are there with a two-word name (e.g. South Africa). (2)

Guidelines

- For the purpose of this assignment, show your code. (5)
- Organize your document into neat sections.
- Your document should not contain any raw R output. Tables should be neatly formatted, values should appear in a written sentence.
- Tables and Figures should have captions.

- Use `ggplot` for all graphics, and `dplyr` for all data management, as far as possible.
- Use the original data files exactly as they are (including names). Once you have read the data into R, you can change variable names, etc. Everything should be reproducible given your code.
- Never use `setwd()` in Rmarkdown / Quarto files.