Global development in 2007:
A reproducible analysis using Gapminder data

Master's thesis

for acquiring the degree of
Master of Science (M.Sc.)

in Business Administration

at the School of Business and Economics
of Humboldt-Universität zu Berlin

submitted by
Max Mustermann
Student no. xxxxxx

First Examiner: « name of first examiner »
Second Examiner: « name of second examiner »

Berlin, May 25, 2025

Table of contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my professors, colleagues, and friends for their support and insights throughout the preparation of this thesis. Special thanks go to [Name], whose guidance was invaluable during the project.

Abstract

This thesis template explores global development patterns using the publicly available Gapminder dataset, with a specific focus on the year 2007. Leveraging a reproducible workflow adapted from the TRR 266 Template, the project presents descriptive statistics by continent and a visual analysis of the relationship between life expectancy, income per capita, and population. The findings reveal clear disparities in health and wealth outcomes across world regions. A comparison with Gapminder's 2021 global development chart highlights persistent global inequalities, while illustrating notable progress in certain regions. The analysis serves as a modern, transparent example of public data storytelling in empirical research.

List of Abbreviations

GDP Gross Domestic Product

TREAT TRR 266 Template for Reproducible Empirical Accounting Research

1 Introduction

This project serves as a template-based example for conducting and documenting reproducible empirical data analysis using publicly available datasets. It leverages the TRR 266 Template for Reproducible Empirical Accounting Research (TREAT) and builds upon the methodological foundations of the Corporate Decision-Making and Quantitative Analysis course, as well as the Accounting Reading Group. To support future empirical thesis projects, this template was developed in adherence to the formal formatting and content guidelines of the Institute of Accounting and Auditing at the School of Business and Economics, Humboldt-Universität zu Berlin. It is not a completed thesis but a demonstrative project hosted in the template repository.

This template project explores global development patterns using the Gapminder dataset, focusing on the year 2007. The project illustrates how reproducible research workflows can be applied by using public and open-source data. Specifically, it presents a summary statistics table by continent and visualizes the relationship between GDP per capita and life expectancy, offering a static snapshot of world development in 2007. These descriptive insights are presented in Section 2.

The results are further contrasted with Gapminder's interactive global development visualizations as of 2021 (Gapminder Foundation 2021). This comparative discussion is provided in Section 3. The project concludes with reflections on transparency, reproducibility, and importance of making research accessible to a broader audience, as discussed in Section 4.

2 Descriptive Analysis of Global Development in 2007

This section explores global development patterns using Gapminder data for the year 2007. It presents a continent-level summary statistics table and a scatter plot showing the relationship between GDP per capita and life expectancy. These outputs serve as a static snapshot of global well-being during that year and demonstrate how transparent and reproducible workflows can be applied to public datasets.

2.1 Research Design Choices and Assumptions

In line with Gapminder Foundation (2021), the analysis uses the Gapminder dataset available via the `gapminder` Python package. This dataset contains country-level information on population, GDP per capita (inflation-adjusted), and life expectancy for various years in five-year intervals. For this project, I focus on the year 2007 - the most recent year available in the dataset at the time.

To ensure clarity and consistency, several assumptions are applied. The analysis is restricted to the year 2007, assuming that cross-sectional variation in life expectancy and GDP per capita during that year sufficiently captures key development patterns. No imputation is performed; only complete observations are included. GDP per capita is used in inflation-adjusted international dollars, as provided by Gapminder Foundation (2021), without applying additional currency normalization. These choices prioritize interpretability and reproducibility, while acknowledging that dynamic trends and structural differences across regions are not captured in this static snapshot. All data originates from the Gapminder Foundation, which curates and harmonizes publicly available indicators to promote global development awareness (Gapminder Foundation 2021).

These assumptions, together with the procedural details in Section 2.2, guide the replication and ensure transparency in design and implementation.

2.2 Replication Steps

This section outlines the modular workflow used to generate the summary table and figure, aligned with Gapminder Foundation (2021).

Step 1: Pulling the Data

The data is loaded from the `gapminder` Python package. The 2007 subset is saved in `.parquet` format for modern data handling and versioning.

Step 2: Data Preparation

The dataset is checked for missing values, which are reported and excluded. All numerical columns are rounded to two decimal places. No transformations, sorting, or new variables are added to preserve the original data structure.

Step 3: Analysis Implementation and Reproduction of Tables and Figure

In the final step, two key outputs are generated to illustrate global development patterns in 2007. First, a summary table is produced, aggregating life expectancy, GDP per capita, and total population figures by continent. This

table provides a concise overview of regional development differences. Second, a scatter plot is created to visualize the relationship between GDP per capita and life expectancy, where each country's bubble size corresponds to its population and color represents its continent.

2.3 Results

This section presents a static snapshot of global development in 2007 using the Gapminder dataset. The figure and table illustrate clear regional disparities in wealth and health outcomes - highlighting, for example, the relatively high life expectancy in Europe and Oceania compared to lower-income regions like Africa.

Table 1 provides aggregated statistics by continent and complements the visual by quantifying central tendencies and demographic scale. It presents average and median values for life expectancy and GDP per capita, along with total population and number of countries represented per continent in 2007. A total of 142 countries are included in the 2007 snapshot.

[Table 1 about here.]

Figure 1 illustrates a scatter plot based on Gapminder's 2007 data, mapping GDP per capita against life expectancy. Each country is represented as a bubble whose size corresponds to its population and whose color indicates its continent. The plot reveals a positive relationship between economic prosperity and life expectancy.

[Figure 1 about here.]

Together, these outputs provide a clear and interpretable overview of global development disparities, offering insights into how income and health outcomes align across continents at a given point in time.

3 Global Development Patterns in 2021

This section compares the 2007 snapshot with Gapminder's 2021 visualization to assess whether key patterns - like the link between GDP per capita and life expectancy - persist and to highlight shifts in global inequality and regional development.

3.1 Comparative Assumptions and Design Choices

This section outlines the assumptions for comparing the 2007 Gapminder dataset to a visualization of global development patterns in 2021.

The assumptions guiding the comparison are as follows: visual inspection is used to assess trends in life expectancy and income per capita; bubble size reflects each country's population, as defined by Gapminder; no new calculations are applied - the 2021 snapshot displays curated values from the Gapminder Foundation (2021).

The relationship between country-level variables and their graphical representation in the visualizations can be summarized as shown in Equation 1:

$$\text{Bubble}_i = f\left(\text{GDP}_i,\ \text{LifeExp}_i,\ \text{Pop}_i,\ \text{Continent}_i\right) \tag{1}$$

3.2 Results

This section compares the 2007 snapshot, rendered from Gapminder data in Python in Section 2, with Gapminder's official visualization of global development in 2021 presented below.

[Figure 2 about here.]

4 Conclusion

This thesis template demonstrates how the TRR 266 framework can be applied to structure a reproducible and transparent empirical accounting study. Using Quarto proved to be a particularly effective and user-friendly solution for structuring and rendering the thesis. It facilitates seamless integration of code, results, and narrative, making it an ideal tool for students at the Institute of Accounting and Auditing at the School of Business and Economics, HU Berlin. Thanks for reading!

References

Gapminder Foundation. 2021. "World Health Chart." https://www.gapminder.org/fw/world-health-chart/.

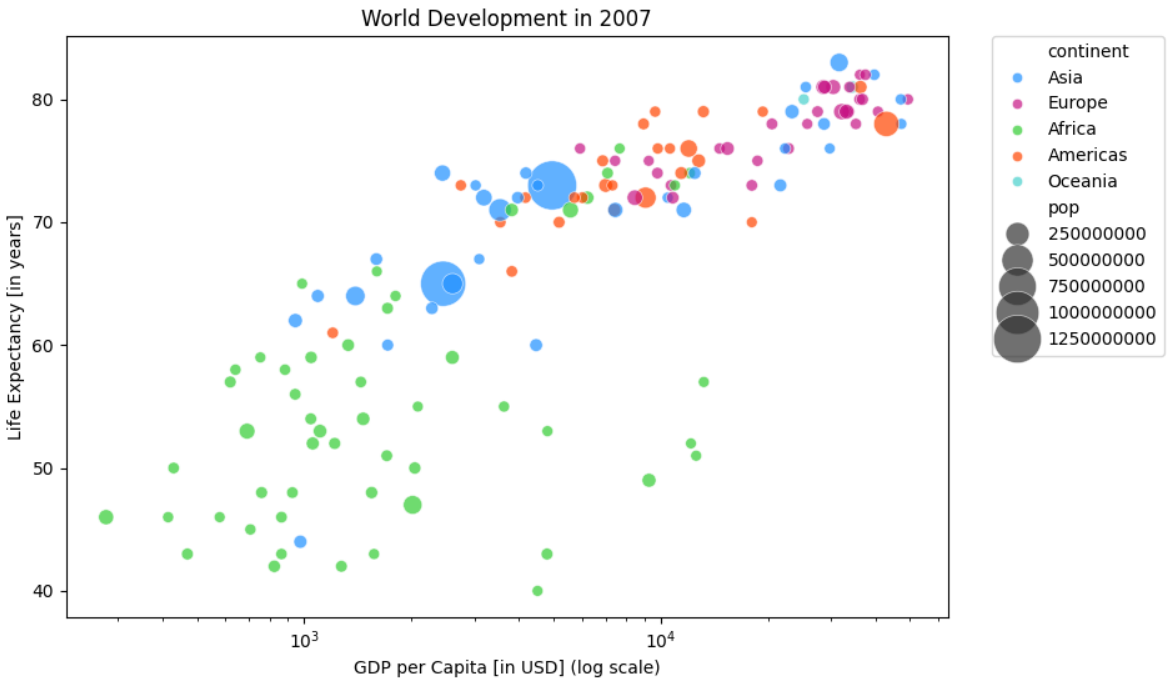

Figure 1: Gapminder 2007: Life Expectancy vs. GDP per Capita

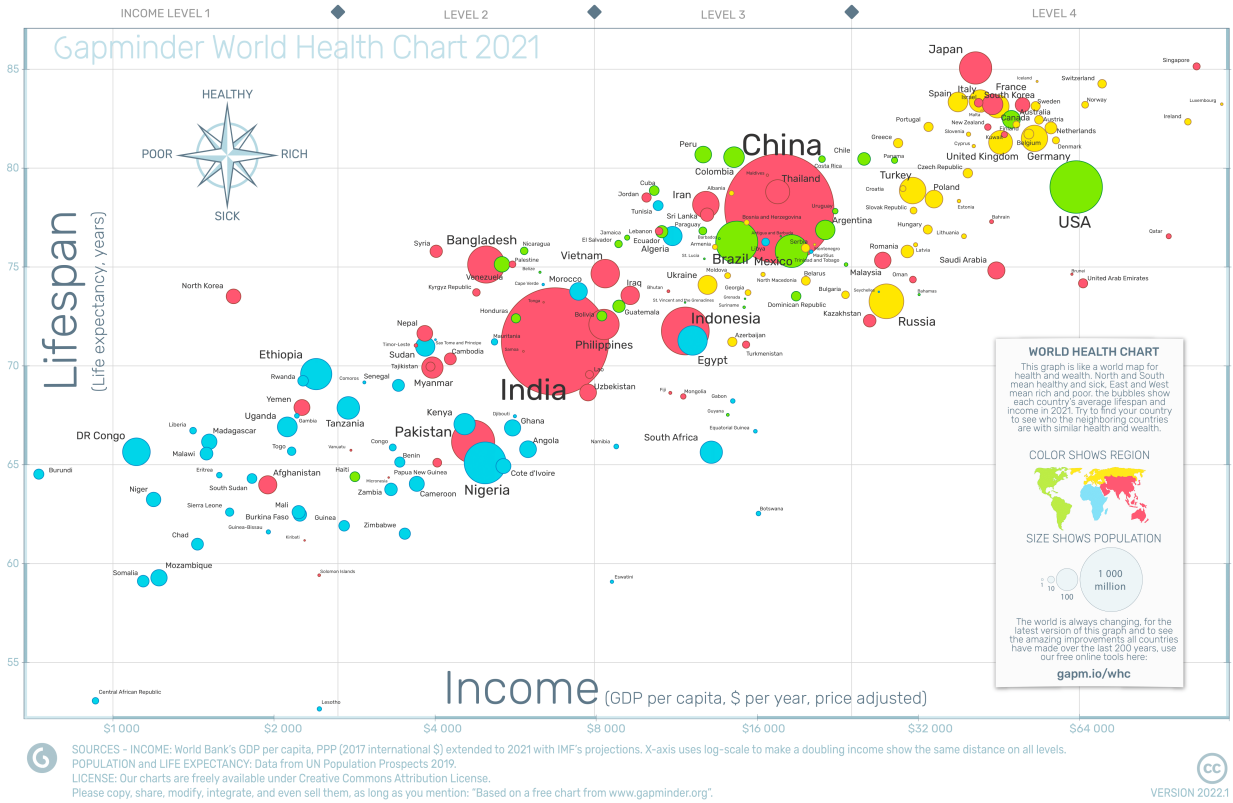

Figure 2: Original World Health Chart 2021 from Gapminder Foundation

Table 1: Summary Statistics by Continent (Gapminder 2007)

continent	Mean_LifeExp	Median_LifeExp	Mean_GDPpc	Median_GDPpc	Total_Pop	Num_Countries
Africa	54.79	53.00	3088.98	1452.00	929539692	52
Americas	73.64	73.00	11003.04	8948.00	898871184	25
Asia	70.82	72.00	12473.00	4471.00	3811953827	33
Europe	77.63	78.50	25054.40	28054.00	586098529	30
Oceania	80.50	80.50	29810.00	29810.00	24549947	2

This table reports the mean and median life expectancy and GDP per capita, along with the total population and number of countries per continent for the year 2007.

Declaration of Academic Honesty

I, Max Mustermann, hereby declare that I have not previously submitted the present work for other examinations. I wrote this work independently. All sources, including sources from the Internet, that I have reproduced in either an unaltered or modified form (particularly sources for texts, graphs, tables and images), have been acknowledged by me as such.

I understand that violations of these principles will result in proceedings regarding deception or attempted deception.

Max Mustermann

Berlin, May 25, 2025