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 23, 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, as well as the Accounting Reading Group course. To support future empirical thesis projects, this template was adapted and developed specifically for students at the School of Business and Economics at Humboldt-Universität zu Berlin—particularly those affiliated with the Institute of Accounting and Auditing and the Finance Group. The template was developed in adherence to the formal formatting and content guidelines of the School of Business and Economics, and in particular, those of the Institute of Accounting and Auditing. 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 beyond traditional financial datasets 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 during 2007. These descriptive results are presented in Section 2.

The results are further contrasted with Gapminder's interactive global development visualizations as of 2021 (Gapminder Foundation 2021), highlighting the potential and limits of static data analysis. This comparative discussion is provided in Section 3. The project concludes with reflections on transparency, reproducibility, and the 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. Results are later contrasted with Gapminder's interactive chart from 2021 in Section 3.

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 exclusively 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. Population is treated as a size indicator in the visualization, assuming that within-year population estimates are comparable across countries. GDP per capita is used in inflation-adjusted international dollars, as provided by Gapminder, 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, ensuring version consistency and avoiding external download links. 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 analysis includes both a summary table by continent and a scatter plot visualizing the relationship between gross domestic product (GDP) per capita and life expectancy. 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.

The summary table (see 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.

[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. The use of a logarithmic x-axis emphasizes differences in income levels across regions while maintaining interpretability across a wide range of values.

[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 Cross-Country Replication

This section generalizes the analysis to a non-U.S. market by replicating Figure 1 using 1972–2023 data from Canada. This involves aligning CRSP/Compustat variables with Worldscope and Datastream equivalents while accounting for differences in reporting standards, market liquidity, and institutional context. Though the core methodology is retained, adapting it to international data demands careful mapping and interpretation.

3.1 Research Design Choices and Assumptions

Since this secition applies a generalization and extension approach to a non-U.S. market, it requires adapting the methodology to alternative databases and market structures. Like in the original study, no analyst forecasts or earnings surprises are used—only stock returns and earnings announcement dates are required. Canada was selected for its comparable quarterly reporting frequency and institutional similarity to the U.S., making it a suitable choice for cross-country replication (**short_short_2025?**). Canadian firms reporting under U.S. GAAP further enhance comparability.

To address database-specific issues not covered in Section 2.1, I make the following adjustments: First, Worldscope/Datastream workflows are kept separate from CRSP/Compustat to prevent interference and facilitate debugging, given the scale of data involved. Second, I exclude fiscal-year-end-based subgroup analysis (Table 2 Panels C–D), as Figure 1 uses the full sample without splitting by fiscal period. Third, I use Worldscope items 5901–5904 to define earnings dates, though limited availability before 1992 may reduce the sample size (**thomson_financial_worldscope_2007?**). Fourth, differences in update frequency between quarterly Worldscope and weekly Datastream can lead to timing mismatches when aligning fundamentals and returns. Fifth, I rely on Datastream's `ret` variable for percentage returns, ignoring bid/ask spreads. Finally, currency effects are ignored, as percentage returns are unaffected by currency denomination.

3.2 Replication Steps

The process is same as in Section 2.

Step 1: Pulling the Data and Managing the Databases

Following (**dai_research_2021?**), I link Worldscope and Datastream using code and infocode. Since Worldscope's `year_` starts in 1980

(**wharton_research_data_services_worldscope_2025?**), Datastream data is also restricted to 1980–2023 to ensure consistent coverage. I extract quarterly earnings announcement dates (items 5901–5904) from Worldscope and daily stock returns from Datastream, focusing on Canadian firms and filtering for common equity (typecode = EQ). This step yields three datasets: 27.9M (Datastream), 77.9K (link table), and 1.94M (Worldscope) observations.

Step 2: Data Preparation

Worldscope is first merged with the link table via code, producing 747,760 rows. This is then joined with Datastream via infocode, resulting in 10.2M observations. I exclude firms missing any earnings date fields, removing 3.6M rows, and drop 6M rows where earnings do not span all four quarters within the same calendar year. To retain more valid windows, event days (-1, 0, 1) are dynamically shifted within ±3 days when needed. After cleaning, I compute 3-day BHRs and extract full-year stock return data from Datastream. All retained infocode entries have around 250 trading days per year, confirming data completeness for annual BHR computation.

Step 3: Analysis Implementation and Reproduction of Tables and Figure

The analysis replicates Table 1 (summary statistics), Table 2 (yearly regressions), and Figure 1 (Abnormal R^2 and slope trends). Using `BHR_3day` and `BHR_Annual`, I run yearly regressions of annual returns on earnings-window returns and compute Abnormal R^2 (adjusted $R^2 - 4.8\%$), handling missing values appropriately.

Specifically, buy-and-hold return over the three-day event window, that measures the stock's reaction only to earnings news, is computed as shown in Equation 1:

$$BHR_{\text{event}} = (1 + R_{t_1})(1 + R_{t_2}) \dots (1 + R_{t_T}) - 1, \quad \text{where } R_t \text{ is the daily return and } T \text{ is the total trading days in a year.} \tag{1}$$

3.3 Results

Comparing the replication results with those from (**ball_how_2008?**) reveals both similarities and discrepancies. The original figure shows a general upward trend in abnormal R^2 , peaking in the early 2000s, whereas the replicated figure captures greater volatility in the later years (2010–2023). Additionally, Ball's figure suggests a smoother long-term dynamic, while the replication exhibits more pronounced fluctuations, particularly post-2010. Panel B in replicated figure exhibits higher volatility and larger coefficient

magnitudes, suggesting increased earnings informativeness post-2006, while Ball's original figure has more stable and lower-magnitude coefficients.

The post-2000 period in the replication shows more inconsistent abnormal R^2 values, possibly reflecting evolving market structures, shifts in disclosure practices, and increased earnings informativeness following regulatory changes (e.g., IFRS adoption, post-SOX adjustments, different financial market structure). The pronounced fluctuations could also be driven by global financial crisis, Canada's higher market concentration, fewer publicly traded firms, and potential regulatory differences, which may affect earnings informativeness and stock return patterns. The variability in slope coefficients further suggests that earnings announcements' impact on stock returns is less stable than in prior decades, potentially due to differences in investor response.

This section compares the static 2007 snapshot with Gapminder's dynamic global development chart Figure 2.

[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 or the Finance Group 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/.

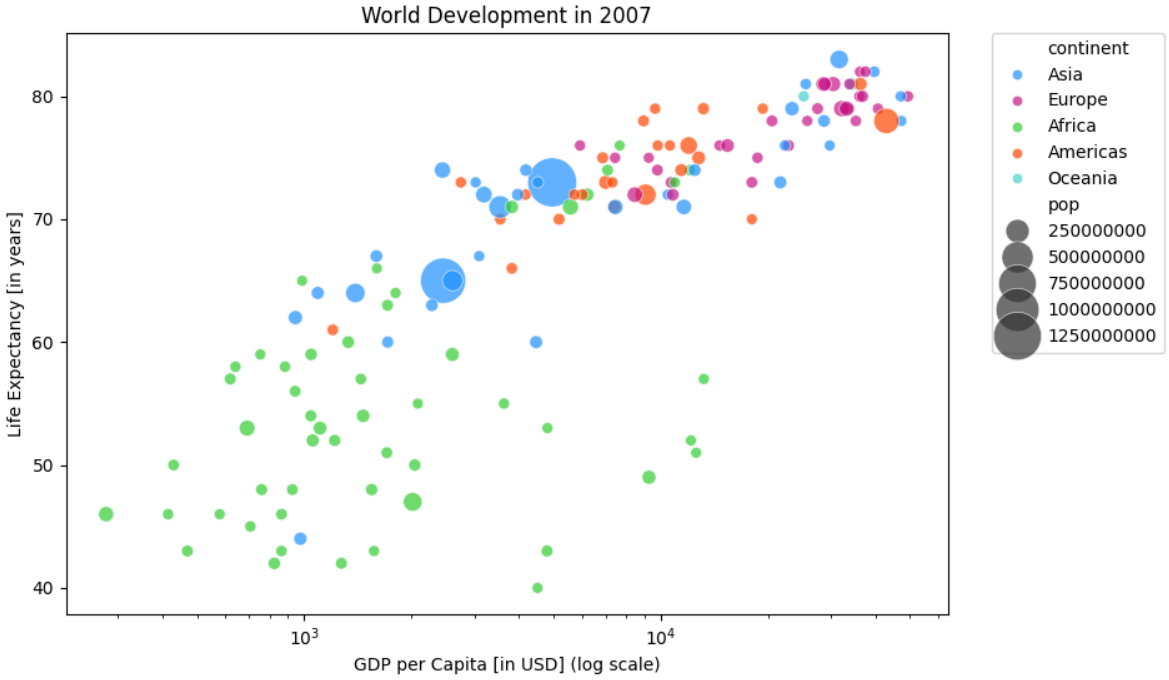

Figure 1: Gapminder 2007: Life Expectancy vs. GDP per Capita

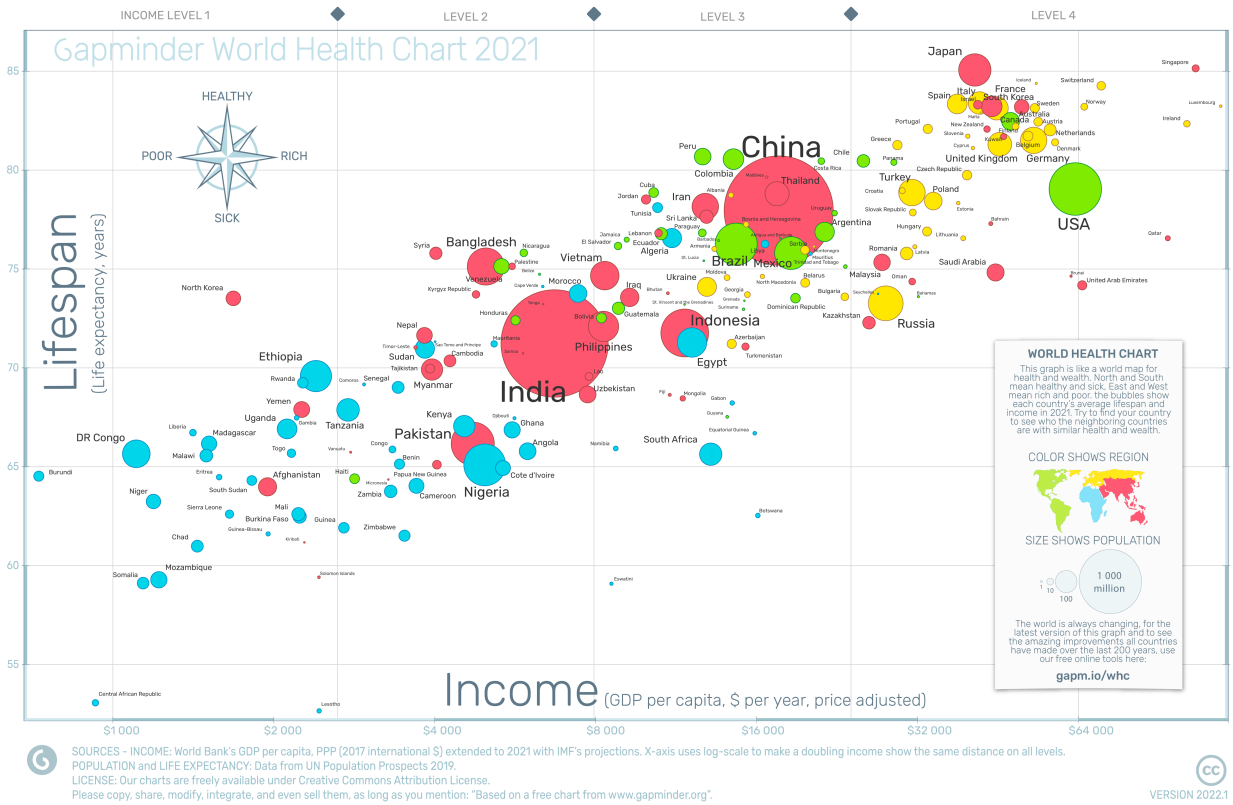

Figure 2: Original World Health Chart 2021 from Gapminder Foundation

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 23, 2025