

SPEC Lab R Resources: Reshaping Data To Make it Tidy – Homework

Jasmine Chu, Claudia Salas Gimenez, Anna Baidar and Ben Graham

Spring 2025

Welcome to the last assignment of Module 3: Data Management II. This homework assignment is designed to help you apply the data tidying and merging skills you've developed through the walkthroughs and group exercises. The goal is to further practice the content covered in this module and prepare you for handling datasets effectively in social science research. By working with real-world data from the World Bank Ed Stats and World Development Indicators databases, you'll gain hands-on experience transforming untidy data into analysis-ready formats, merging datasets, and assigning meaningful variable names.

Save your responses in your personal subfolder in the 412_413 `shared` Google Drive folder. The R script should be titled `HW_DM2_[YOUR_INITIALS]`. You can also save a copy of your R script to your own computer for future reference.

Initial Setup

Before starting, ensure you've set up your working directory and installed the necessary libraries. Also, download the `world_bank_education.csv` dataset from the World Bank Ed Stats datasets, which contains Barro and Lee's measures for estimates for the population between 15 and 19 year-old that have completed secondary school.

To gain a deeper understanding of the datasets and their variables, refer to the [World Development Indicators data catalog](#) to better understand the datasets and their variables. Keep in mind that since these datasets have been modified for this walkthrough, the variable names may differ slightly from those in the original codebooks. [Barro-Lee website](#) to view the list of variables, countries, data sources and brief description of the methodology, as well as access the latest updated version of the Barro-Lee dataset.

The goal of this assignment is to tidy the `world_bank_data_literacy_rates.csv` data into a country-year format.

Exercise 1: Reshaping Data

Exercise 1.1: Rename Variables

Before reshaping the dataset, start by renaming the columns referring to the years 1970 through 2020 to their respective year numbers.

Exercise 1.2: Reshape and Clean the Data

Next, transform the dataset into a tidy format where each row represents a unique combination of country and year.

Helpful Hint: Remember to use backticks (```) around variable names that contain spaces to ensure they are properly recognized in your code.

Exercise 1.3: Rename the New Variables

Lastly, assign concise, descriptive names to the variables.

Exercise 2: Include Additional Variables

Now that you've reshaped and cleaned educational data, let's dive deeper by adding more dimensions to the analysis. In this exercise, we'll expand the dataset by including data on literacy rates for population between 15 and 24 from the World Development Indicators.

Exercise 2.1: Load and Rename Variables

Let's start by loading the `world_bank_data_literacy_rates.csv` data from the Training Data Spring 2024 folder, and rename the columns for the years 1970 through 2020 to their respective year numbers.

Exercise 2.2: Reshape and Clean the Data

As done with the `world_bank_education.csv` dataset, transform the literacy rate dataset into tidy format, where each row represents a unique country-year combination.

Helpful Hint: Remember that if variable names consist of multiple words separated by spaces, you'll need to enclose them in backticks (`` ``) to ensure they are recognized correctly in your code.

Exercise 2.3: Rename the New Variables

Also, provide clear and concise names for literacy rate-related variables.

Exercise 3: Merging Datasets

Now that both datasets are clean and tidy, it's time to combine both datasets. Merge the education and literacy data, keeping all available data from both datasets.

Note: The standard practice in SPEC Lab is to merge datasets using Gleditsch-Ward Country Codes and year for precise matching. However, for this assignment, since both datasets originate from the World Bank, merging by country name and year is sufficient.

Exercise 4: Visualizing the Relationship Between Literacy Rates and Educational Attainment

Use the merged dataset to create a scatterplot of the relationship between the literacy rates and educational attainment in a single year.

Bonus Exercise 1

As bonus question, plot the over time trends in the global average of literacy rates and educational attainment.

Bonus Exercise 2

Although in Bonus Exercise 1 we plotted overall trends in literacy rates and educational attainment, our final dataset contains many missing values. This means the patterns we observe might primarily reflect only those countries that did report literacy rates and educational attainment. In this exercise, plot the over-time trends in the global average of literacy rates and educational attainment for countries with non-missing values. What types of countries are most likely to report their literacy rate and educational attainment data?