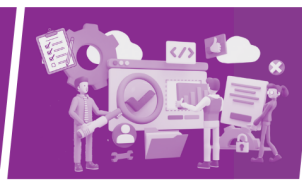
Fall 2023

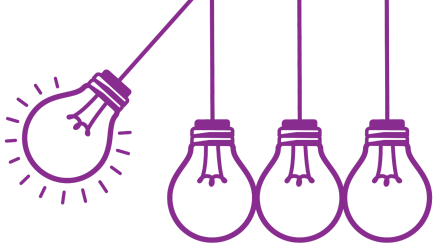
DIME Analytics

REPRODUCIBLE RESEARCH FUNDAMENTALS



THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT



Tidying data
R (primary data) exercise

Reproducible Research Fundamentals

September 26, 2023

Development Impact Evaluation (DIME)
The World Bank

- During the training, find all materials in our shared OneDrive: https://bit.ly/rrf23-materials

WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT

Overview

Preamble

- **Slides:** You will find these slides in `Course_Materials/Labs/Primary/R`
- **Data:** The hands-on sessions will use the data from LWH (Land husbandry, Water harvesting, and Hillside irrigation) project, an impact evaluation of agricultural development in Rwanda.
  - Data shared in OneDrive folder: `Course_Materials/Labs/Primary/R/data`
  - Case study and questionnaire: `Course_Materials/Labs/Primary`
- **Templates:** You can create your code from scratch or use the template scripts: `Course_Materials/Labs/Primary/R/scripts`

Exercises

Exercise 1

Exercise 1: Explore the data

1. Open the template script for tidying data
2. Load the dataset `LWH_FUP2.dta`
3. Explore the data and the documentation:
    - What is the unit of observation in the dataset?
    - Does the data have a unique ID?
    - Do all the variables in the dataset have the same unit of observation?
    - Is there more than one unit of observation in this dataset?

Exercise 1

Useful commands:

- `read_stata()` from the library `haven` to read Stata files into R dataframes
- To check if a dataset or column(s) have unique values:
  - `n_distinct()` from `dplyr`
  - `nrow()`
  - `select()` and the selection helper `any_of()` might be useful here

Exercise 2

Exercise 2: Fix duplicates

1. Remove any duplicated observations, either for cases when the entire observation is duplicated or when the ID variables are duplicated
2. Add a documentation Word or text file in your documentation folder explaining which duplicate you are dropping and why you selected that observation

Exercise 2

Useful commands:

- `filter()` from `dplyr` to drop or keep observations
- `group_by()` might be useful to group the dataframe by a number of variables and count duplicates by the grouped variables with `n()`

Exercise 3

Exercise 3: Create tidy datasets

1. Split the untidy dataset into tidy datasets for each unit of observation used in any of the variables
   - How many datasets will you create?
   - What is the unit of observation of each dataset?
2. Save each tidy dataset into a file

Exercise 3

Useful commands:

- `select()` from `dplyr`
- `any_of()` will be a useful selection helper for this exercise (from `dplyr` or `tidyselect`
- `pivot_longer()` from `tidyr` to reshape in long format
- `mutate()` and `recode_factor()` might help you recode the dataframes after reshaping

Discuss - How can tidyness help you?

Discuss

- Are there any next steps in data work that have been made easier after tidying the dataset?
- What indicators are easier to construct after tidying the data?

Thanks! Gracias!