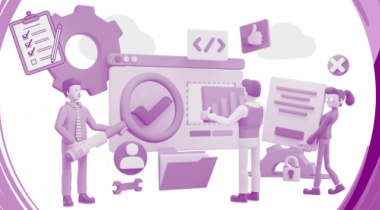
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

WORLD BANK GROUP
INSTITUTE FOR ECONOMIC DEVELOPMENT

Reproducible Research Repository

Processing Data for Reproducible Analysis



Overview

Overview - Data

- **Primary Data**: The data for this analysis comes from an impact evaluation of Community-Based Conditional Cash Transfers (CCT) in Tanzania. This program aimed to replicate successful CCT models by involving communities in beneficiary targeting and payment distribution to improve outcomes for the poor. There are two datasets:
  - TZA_CCT_baseline.csv
  - treat_status.csv
- **Templates**: You can code from scratch or you can use the template scripts from the link.

Overview - Lab sessions

- These lab sessions are exercises of the data workflow in reproducible research.
- All exercises assume previous knowledge of R.
- The exercises suggest which commands or libraries to use for solving them.
- The template scripts offer additional information. You're free to follow the given suggestions or solve the exercises in any other way.

Exercises

Exercise 1

Exercise 1: Explore the data

1. In Rstudio, open the script for data processing (`01-processing-data.R`).
2. Load the dataset `TZA_CCT_baseline.csv`
3. Explore the data:
   - 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 2

Exercise 2: Fix duplicates

Remove any duplicated observations, either for cases when the entire observation is duplicated or when the household ID variable is duplicated

Tidying Data

Tidying Data

In this exercise, you will work with a single dataset that is already in a tidy format for simplicity. However, for future projects, it's important to consider the following:

- How many data frames will you need to create?
- What will be the unit of observation for each data frame?
- Will you need to reshape any of the data frames?

Data Cleaning

Data cleaning tasks

Some of the data cleaning tasks are:

- Assign variables to the correct data types
- Fix missing values
- Explore ”other” variables and encode them if needed
- Drop variables that will not be required
- Check for outliers
- Check that all variables have labels and value labels

Exercise 3

Exercise 3: Clean the data

1. For the households dataframe, perform the following data cleaning tasks
   - Replace numeric values representing missing data (-88) with missings.
   - Extend the values in `crop` with the two most used categories of `crop_other`
   - Find out if any numeric variable has outliers.

Exercise 4

Exercise 4: Saving

Save your dataframes into the Intemediate folder.

Final notes: Documentation

- We haven't created documentation in these exercises, but it should also be an output of data processing
- The documentation should include:
  - A summary of the results of data exploration, data cleaning tasks, and documentation of any decision made in data processing. The deletion of observations with gender equal to missing should be justified there, for example.
  - A metadata table listing each variable, their labels, the number of missing values they have, and some summary statistics.

THANK YOU

WORLD BANK GROUP
INSTITUTE FOR ECONOMIC DEVELOPMENT

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

Reproducible Research Repository