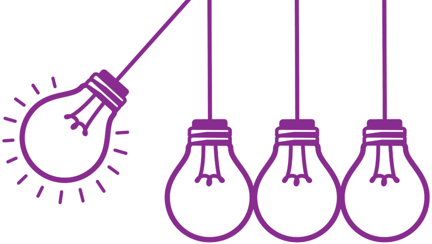
Fall 2023

DIME Analytics

REPRODUCIBLE RESEARCH FUNDAMENTALS

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

i2i
DIME
TRANSFORM DEVELOPMENT



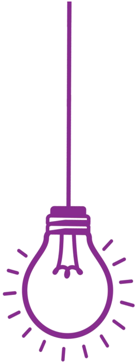
Tidying data - Hands-on
Track 1 - Stata (Primary Data)

Reproducible Research Fundamentals

September 26, 2023

Development Impact Evaluation (DIME)
The World Bank

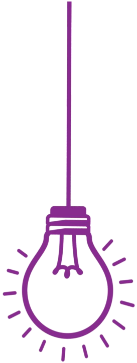
- During the training, find all materials in our shared OneDrive: here



Overview

Overview

- **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/Stata/data`
    - Associated Case Study and Questionnaire: `Course_Materials/Labs/Primary`
- **Templates**: You can code from scratch or you can use the template do-files: `Course_Materials/Labs/Primary/Stata/scripts`



Exercises

Exercise 1: Explore dataset

1. Open the template do-file for tidying data
2. Load the dataset `LWH_FUP2.dta`
3. Explore the data:
    - What is the unit of observation in the dataset?
    - Does the data have a unique ID?
        - Commands for testing that a variable is uniquely and fully identifying: `isid` or `codebook`
    - 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: Fix duplicates

We will fix duplicates by using `ieduplicates` command from `iefieldkit` package. You can install `iefieldkit` by typing:

```
ssc install iefieldkit, replace
```

`ieduplicates` identifies duplicates in ID variable and exports them in an Excel file that can be used to correct the duplicates.

Exercise 2: Fix duplicates using `ieduplicates`

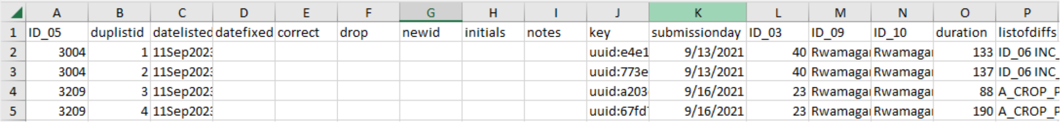
- Use the `ieduplicates` command to identify the duplicates in the Household ID (`ID_05`). Export them to an excel file that can be used to correct the duplicates.

```
    ieduplicates    idvarname ///
                    using ”path/to/duplicates_report.xlsx”, ///
                    uniquevars(varlist) ///
                    keepvars(varlist) ///
                    force
```

- Update the folder path so that the excel file is exported to the right location
- The unique ID in the dataset is `key`
- Input the list of variables that you would want to be included in your report (enumerator ID, location identifiers)

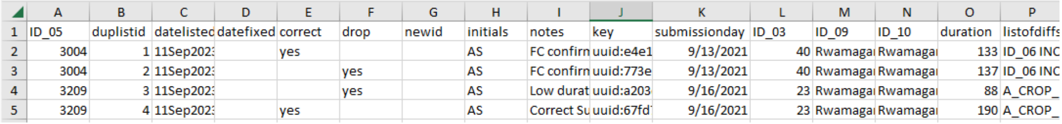
Fix duplicates: Output and corrections

1. After running the code, open the exported excel file:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ID_05	duplistid	datelisted	datefixed	correct	drop	newid	initials	notes	key	submissionday	ID_03	ID_09	ID_10	duration	listofdiffs
2	3004	1	11Sep202							uuid:e4e1	9/13/2021	40	Rwamagai	Rwamagai	133	ID_06 INC_
3	3004	2	11Sep202							uuid:773e	9/13/2021	40	Rwamagai	Rwamagai	137	ID_06 INC_
4	3209	3	11Sep202							uuid:a203	9/16/2021	23	Rwamagai	Rwamagai	88	A_CROP_F
5	3209	4	11Sep202							uuid:67fd	9/16/2021	23	Rwamagai	Rwamagai	190	A_CROP_F

2. Fix the duplicates by editing the fields in the excel:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ID_05	duplistid	datelisted	datefixed	correct	drop	newid	initials	notes	key	submissionday	ID_03	ID_09	ID_10	duration	listofdiffs
2	3004	1	11Sep202		yes			AS	FC confirn	uuid:e4e1	9/13/2021	40	Rwamagai	Rwamagai	133	ID_06 INC
3	3004	2	11Sep202			yes		AS	FC confirn	uuid:773e	9/13/2021	40	Rwamagai	Rwamagai	137	ID_06 INC
4	3209	3	11Sep202			yes		AS	Low durat	uuid:a203	9/16/2021	23	Rwamagai	Rwamagai	88	A_CROP_
5	3209	4	11Sep202		yes			AS	Correct Su	uuid:67fd	9/16/2021	23	Rwamagai	Rwamagai	190	A_CROP_

3. Save the excel and run the code again. The resulting data will be unique in Household ID.

Exercise 3: Create tidy datasets

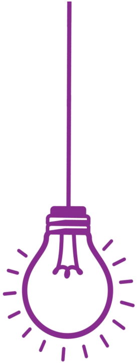
1. Split the untidy dataset into tidy datasets for each unit of observation used in any of the variables
    - How many tidy datasets can be created?

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 tidy datasets can be created?
2. Use `reshape` to make the data tidy where necessary
3. What is the unit of observation for each new tidy dataset?
4. Save the tidy datasets

Discuss - How can tidying help you?

1. Are there any next steps that have been made easier after tidying the datasets?
2. What indicators are easier to construct after tidying the dataset?



The End