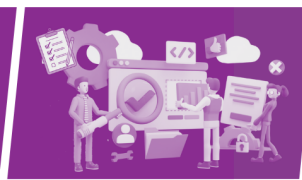
Fall 2023

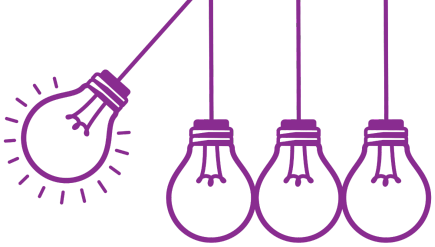
DIME Analytics

REPRODUCIBLE RESEARCH FUNDAMENTALS



THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT



Data Cleaning - Hands-on
Track 1 - Stata (Primary Data)

Reproducible Research Fundamentals

September 27, 2023

Development Impact Evaluation (DIME)
The World Bank

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

WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT

Overview

You will start with the following tidy datasets created in the last exercise and work through typical data cleaning tasks to create clean datasets for each of them:

- `LWH_FUP2_households.dta`
- `LWH_FUP2_assets.dta`
- `LWH_FUP2_plot.dta`

For this session, you can use the template do-file provided.

Data cleaning

What are some of the data cleaning tasks?

Data cleaning tasks

Some of the data cleaning tasks:

- Make sure all the variables have the correct data type
- Fix extended missing values
- Check that all variables have labels and value labels
- Explore “other” variables and encode them if needed
- Drop variables from the survey that are not required anymore
- Explore data to identify outliers

Data cleaning outputs

What are some of outputs created after cleaning the data?

Data cleaning outputs

What are some of outputs created after cleaning the data?

1. The cleaned dataset
2. Documentaion of data cleaning tasks
3. Metadata that stores information like:
    - The definition of each variable or corresponding survey question
    - The number of missing observations in each variable
    - Summary statistics
    - Any field notes or corrections made to each variable

Data cleaning in Stata

Commands in Stata for data cleaning

Here are some commands that can be used while cleaning data:

- `ds, has(type typelist)`: Checks the type of the variables (string/numeric)
- `destring` or `tostring`: Converts string variables to numeric variables and vice versa
- `encode` or `decode`: Encodes string into numeric variable or vice versa
- `recode`: Recodes categorical variables
- `label variable`, `label define`, `label value`, `label dir`, `label list`, `labelbook`: Manipulates labels

`iefieldkit` **and** `ietoolkit`

Using `iefieldkit` to annotate the data set

- The `iecodebook` command (part of `iefieldkit` package) helps you perform most of the tasks described above (with the exception of encoding)
- The command outputs (in Excel) a list of all variables in the data set and their labels, and applies changes to them so the process is simplified
- The Excel report is used to document the modifications made to the data set while cleaning

	A	B	C	D	E	F	G	H	I
1	name	label	type	choices	name:current	label:current	type:current	choices:current	recode:current
2	survey	(Ignore this placeholder, but do not delete it. Thanks!)	float	yesno					
3	dist	District ID		.	dist	Esta comunidade é de qual distrito?	byte	dist	
4	comid	Community ID		.	comid	Qual é esta comunidade?	int	comid	
5	hhid	Household ID			hhid	Introduze o ID do agregado familiar:	long		
6					hhdurablesq	Na sua casa principal, o seu agregado familiar tem...	byte	hhdurablesq	
7	oillamp	Household owns an oillamp		yesno	oillamp	[2.01] Um candeeiro de petróleo?	byte	oillamp	
8	radio	Household owns a radio		yesno	radio	[2.02] Um rádio?	byte	radio	
9	bicycle	Household owns a bycicle		yesno	bicycle	[2.03] Uma bicicleta?	byte	bicycle	
10	latrine	Household has a latrine		yesno	latrine	[2.04] Uma latrina?	byte	latrine	
11	table	Household owns a table		yesno	table	[2.05] Uma mesa?	byte	table	
12	cellphone	Household owns a cellphone		yesno	cellphone	[2.06] Um celular / telemóvel?	byte	cellphone	
13	solar	Household owns a solar panel		yesno	solar	[2.07] Um painel solar?	byte	solar	
14	motorbike	Household owns a motorbike		yesno	motorbike	[2.08] Uma motocicleta / motorizada?	byte	motorbike	
15	tv	Household owns a tv		yesno	tv	[2.09] Uma televisão?	byte	tv	

iecodebook syntax

1. Create a codebook template:

```
iecodebook template using ///
                          "path/codebook.xlsx", ///
                           replace
```

2. Open the excel file and edit it to change variable labels, add value labels, and recode missing values.
3. Save the excel file and then apply the changes:

```
iecodebook apply using ///
                       "path/codebook.xlsx" ///
```

Using `iesave` to save metadata report

- The `iesave` command (part of `ietoolkit` package) automates best practices such as exporting metadata, compressing the data, and testing ID variables
- The command applies best practices before saving the data and also outputs (in .csv or .md) a metadata report that contains information on the ID variable(s), the number of observations, the number of variables, and summary statistics.

Number of observations:	74									
Number of variables:	16									
ID variable(s):	make									
Data signature:	74:16(110350):932916212:1889387683									
Last saved by:	wb501238									
Last saved at:	8/2/2022 11:39									
Variable type: String										
Name	Label	Type	Complete	Number of levels						
make	Make and Model	str17	74	74						
Variable type: Continuous										
Name	Label	Type	Complete	Mean	Std Dev	p0	p25	p50	p75	p100
displacement	Displacement (cu. in.)	int	74	197.3	91.84	79	119	196	250	425
gear_ratio	Gear Ratio	float	74	3.015	0.4563	2.19	2.73	2.955	3.37	3.89
headroom	Headroom (in.)	float	74	2.993	0.846	1.5	2.5	3	3.5	5
length	Length (in.)	int	74	187.9	22.27	142	170	192.5	204	233

iesave syntax

Install ietoolkit by typing:

```
ssc install iefieldkit, replace
```

Use iesave to save the data and export a report:

```
1        iesave   "path/to/data1.dta", ///
2                 idvars(idvariable)  version(version_number) replace ///
3                 report(path("path/to/report.csv") replace)
```

Exercises

Exercises

For each of the tidy datasets mentioned earlier, perform the following exercises:

1. **Data cleaning:**
    - Make sure all the variables have the correct data type
    - Fix extended missing values
    - Check that all variables have labels and value labels
    - Explore “other” variables and encode them if needed
    - Drop variables from the survey that are not required anymore
    - Explore data to identify outliers
2. **Documenting data cleaning and consistency:**
    - Document all the data cleaning tasks and the changes you applied to the dataset
    - Create or export a codebook or report of your cleaned dataset

The End