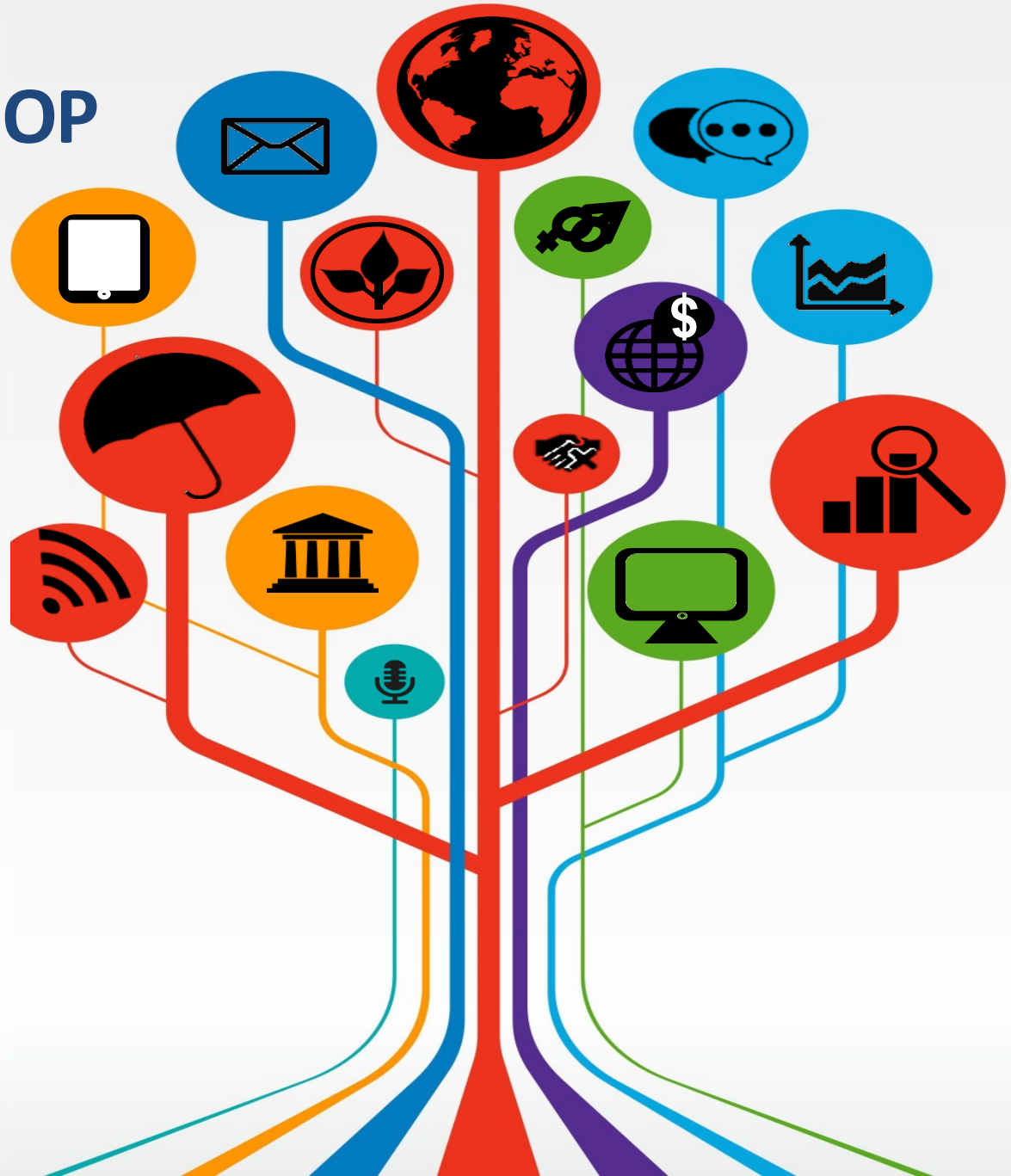
FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC



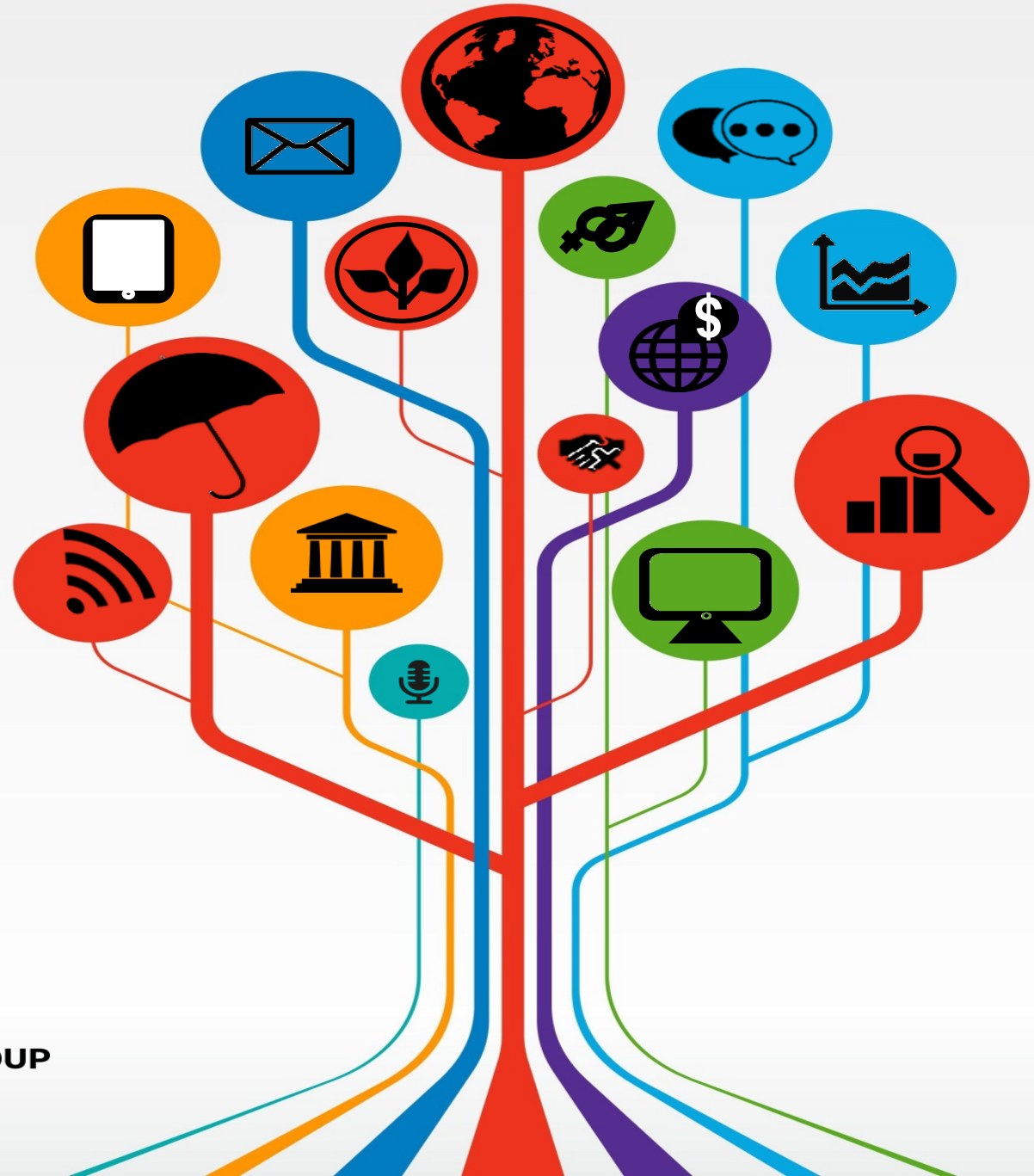
i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Lab 2 – Track 1 - Coding for Reproducible Research

Michael Orevba & Kristoffer Bjärkefur

19 June 2018



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Excel vs. Stata (or R, python etc.)

Can I use Excel?

The main reason we use Stata (or R)

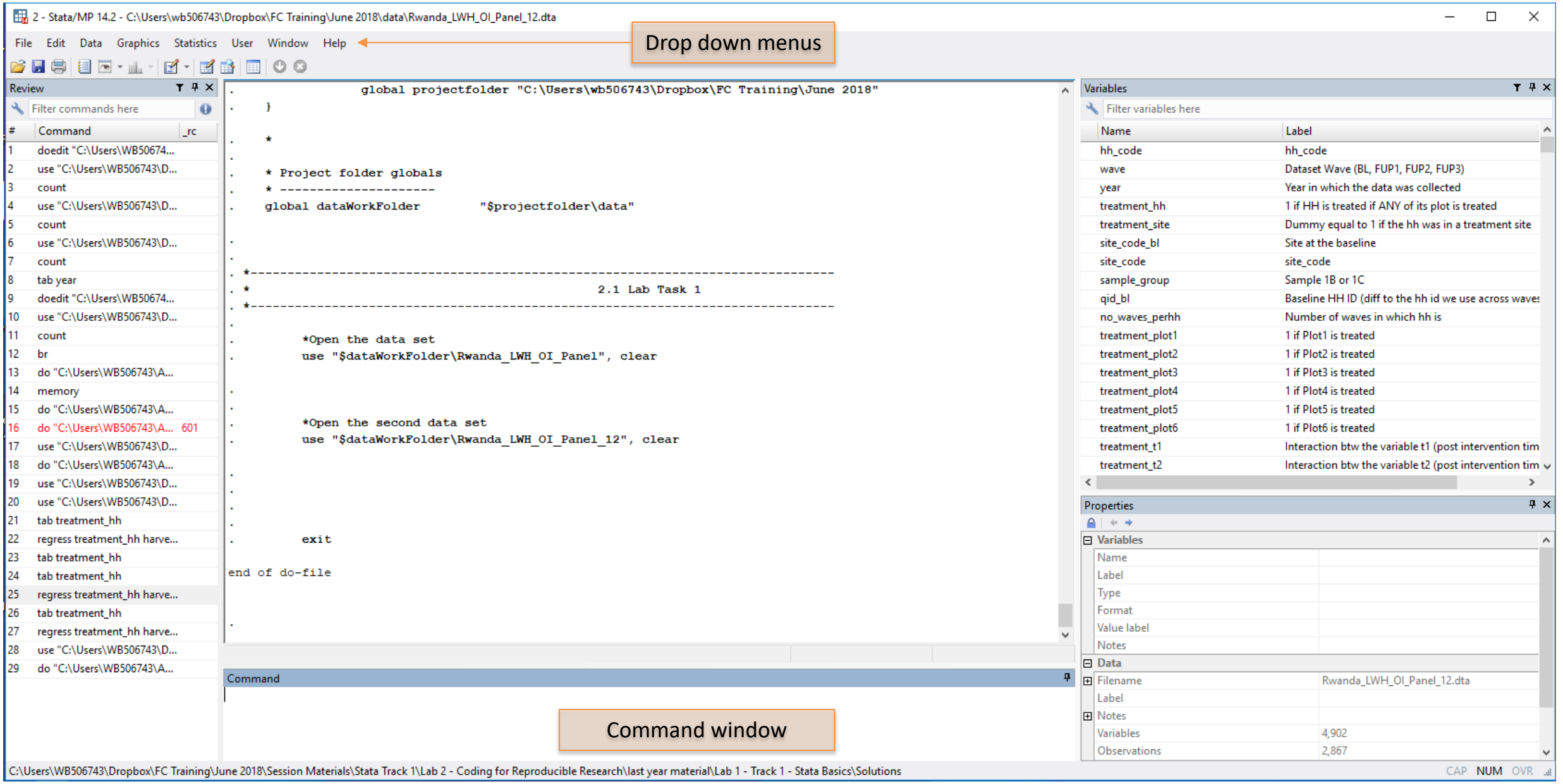
- In Excel you make changes directly to the data and save new versions of the data set

- In Stata you make changes to the instructions on how to get from the raw data to the final analysis and save new versions of the instructions

What's the fuss about do-files?

- It’s through the do-file you communicate your work to other members in your team, both current and future
- Think of the do-files as instructions on how to get from raw data to final report
- Also, for a simple task you can enter commands manually, but for more complex tasks you need to write a recipe, or a list of instructions

How to open up a data set in Stata


2 - Stata/MP 14.2 - C:\Users\wb506743\Dropbox\FC Training\June 2018\data\Rwanda_LWH_OI_Panel_12.dta
File Edit Data Graphics Statistics User Window Help
Drop down menus
Review
Filter commands here
Command _rc
1 doedit "C:\Users\WB50674...
2 use "C:\Users\WB506743\D...
3 count
4 use "C:\Users\WB506743\D...
5 count
6 use "C:\Users\WB506743\D...
7 count
8 tab year
9 doedit "C:\Users\WB50674...
10 use "C:\Users\WB506743\D...
11 count
12 br
13 do "C:\Users\WB506743\A...
14 memory
15 do "C:\Users\WB506743\A...
16 do "C:\Users\WB506743\A... 601
17 use "C:\Users\WB506743\D...
18 do "C:\Users\WB506743\A...
19 use "C:\Users\WB506743\D...
20 use "C:\Users\WB506743\D...
21 tab treatment_hh
22 regress treatment_hh harve...
23 tab treatment_hh
24 tab treatment_hh
25 regress treatment_hh harve...
26 tab treatment_hh
27 regress treatment_hh harve...
28 use "C:\Users\WB506743\D...
29 do "C:\Users\WB506743\A...
. global projectfolder "C:\Users\wb506743\Dropbox\FC Training\June 2018"
. }
. *
. * Project folder globals
. * ---------------------
. global dataWorkFolder "$projectfolder\data"
. *---
. * 2.1 Lab Task 1
. *---
. *Open the data set
. use "$dataWorkFolder\Rwanda_LWH_OI_Panel", clear
. *Open the second data set
. use "$dataWorkFolder\Rwanda_LWH_OI_Panel_12", clear
. exit
end of do-file
Command
Command window
Variables
Filter variables here
Name Label
hh_code hh_code
wave Dataset Wave (BL, FUP1, FUP2, FUP3)
year Year in which the data was collected
treatment_hh 1 if HH is treated if ANY of its plot is treated
treatment_site Dummy equal to 1 if the hh was in a treatment site
site_code_bl Site at the baseline
site_code site_code
sample_group Sample 1B or 1C
qid_bl Baseline HH ID (diff to the hh id we use across waves
no_waves_perhh Number of waves in which hh is
treatment_plot1 1 if Plot1 is treated
treatment_plot2 1 if Plot2 is treated
treatment_plot3 1 if Plot3 is treated
treatment_plot4 1 if Plot4 is treated
treatment_plot5 1 if Plot5 is treated
treatment_plot6 1 if Plot6 is treated
treatment_t1 Interaction btw the variable t1 (post intervention tim
treatment_t2 Interaction btw the variable t2 (post intervention tim
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Filename Rwanda_LWH_OI_Panel_12.dta
Label
Notes
Variables 4,902
Observations 2,867
C:\Users\WB506743\Dropbox\FC Training\June 2018\Session Materials\Stata Track 1\Lab 2 - Coding for Reproducible Research\last year material\Lab 1 - Track 1 - Stata Basics\Solutions
CAP NUM OVR

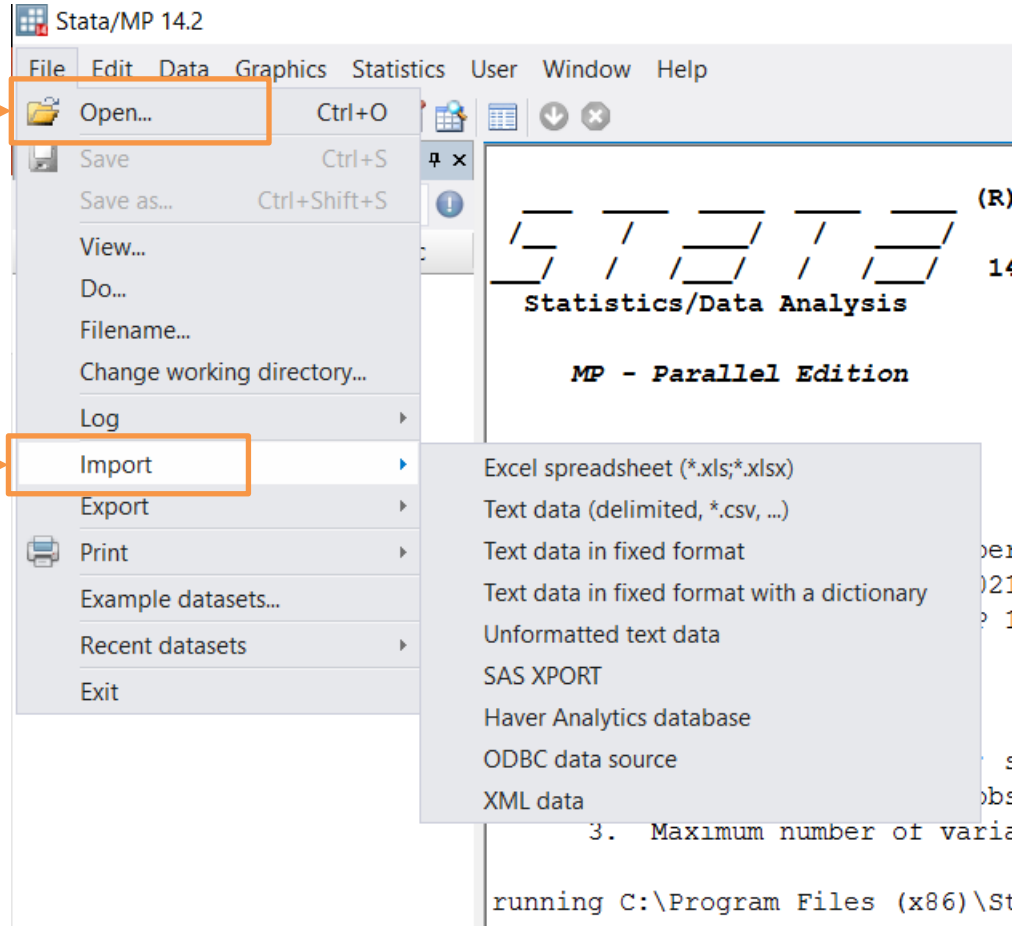
Three ways to tell Stata what to do

- Drop-down menus
  - An easy place to start but quickly becomes inefficient
- Command window
  - Faster than menus but require that you are familiar with the command
- Do-file
  - Use menus and command window to figure out what you need to write, then copy to a do file
  - The only feasible way to run long instructions

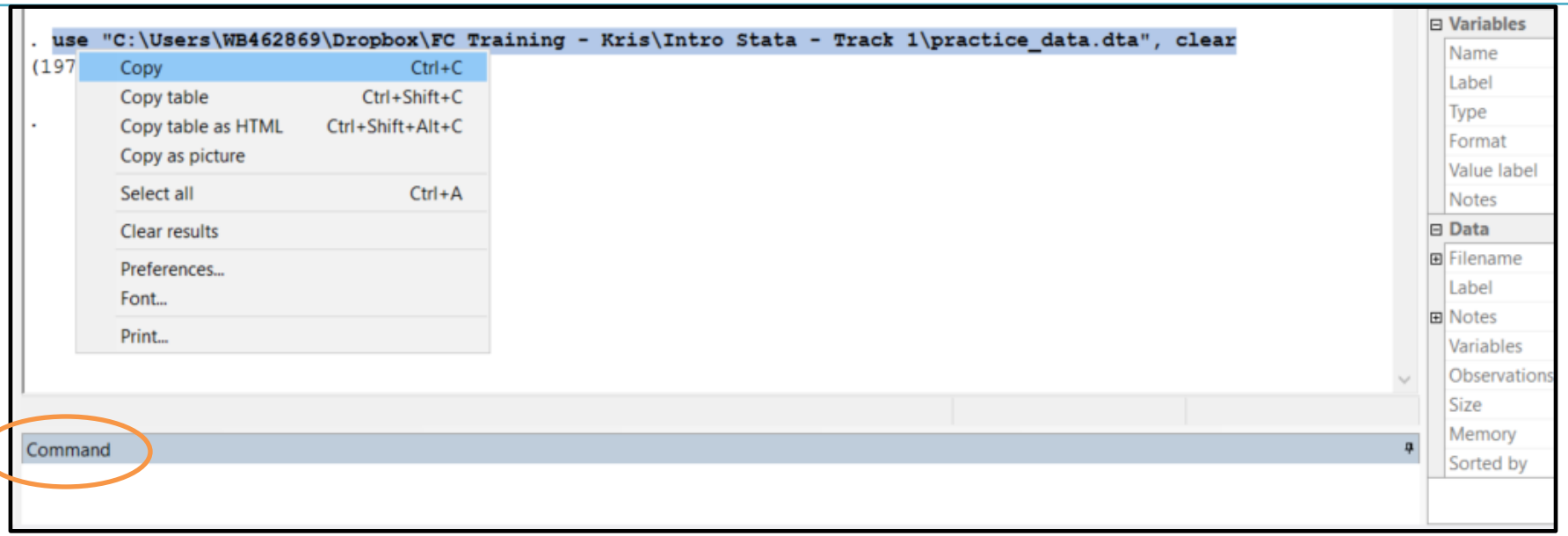
Open a dataset - menus

Open data sets that are already in Stata format here (.dta)

Open all other types of data sets (.csv, .xslx etc.) here



Open a dataset – command window



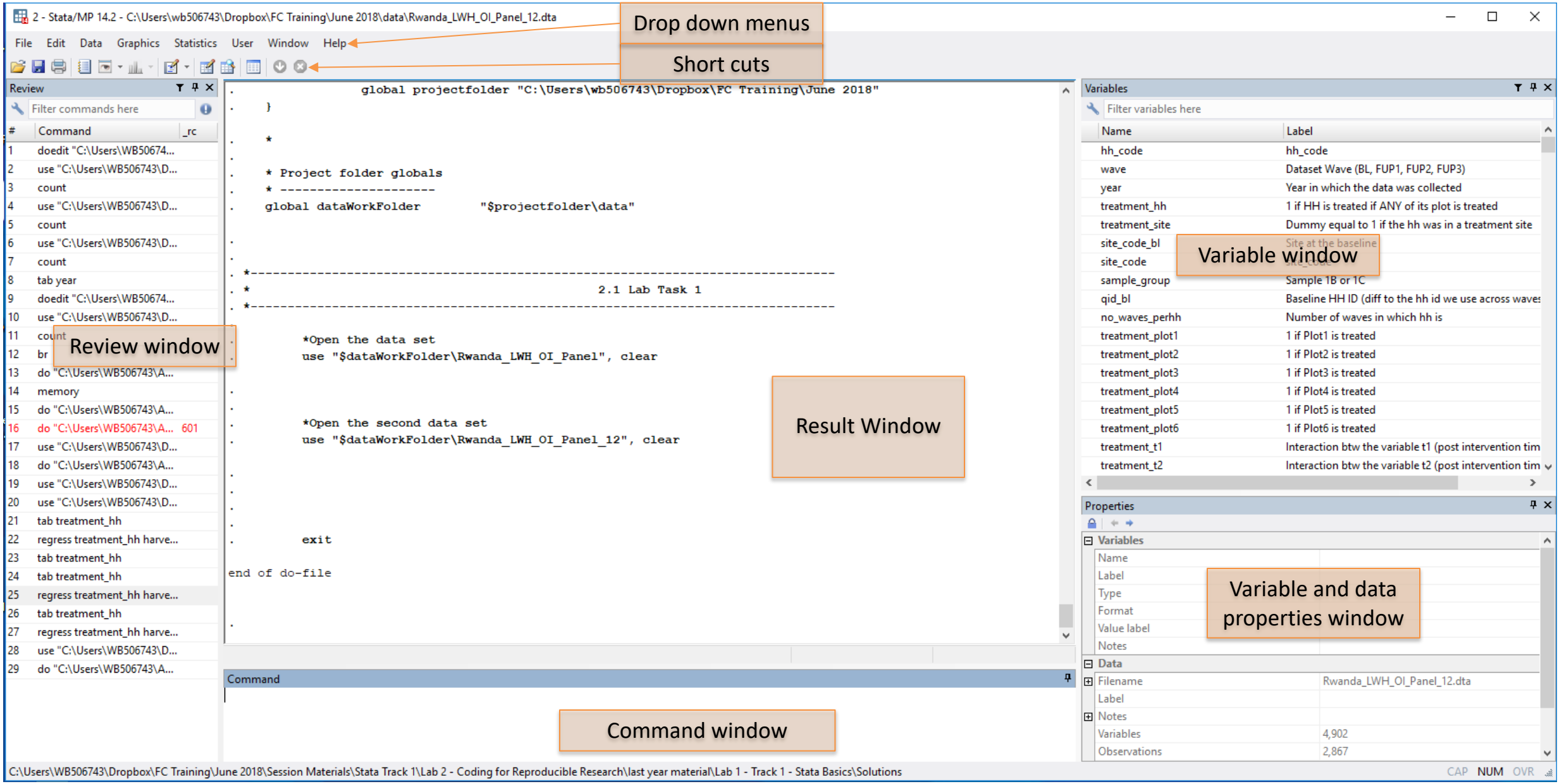
- When you use the menus, Stata produces the code for that action
- Highlight, right-click and copy the code. Paste the code in the command window. And hit enter.

Lab Task 1

1. Open Stata and then open data set *endline_data_raw.dta* using the menu: File -> Open. Navigate to where you saved the material for this lab. Select the data set and click *Open*.
2. Browse to check that you have data: Data -> Data Editor -> Data Editor Browse
3. Copy the code Stata generated for you and paste it in the command window. Hit enter and then browse the data again.
4. Copy and paste the code again. But this time, change the code so that you open *panel_data.dta* instead. You do not have to use the menus again, just change the filename. Browse the data!
5. We will introduce do-files very soon

An introduction to Stata:

Stata interface (what are all those windows)


2 - Stata/MP 14.2 - C:\Users\wb506743\Dropbox\FC Training\June 2018\data\Rwanda_LWH_OI_Panel_12.dta
Drop down menus
File Edit Data Graphics Statistics User Window Help
Short cuts
Review
Filter commands here
Command _rc
1 doedit "C:\Users\WB50674...
2 use "C:\Users\WB506743\D...
3 count
4 use "C:\Users\WB506743\D...
5 count
6 use "C:\Users\WB506743\D...
7 count
8 tab year
9 doedit "C:\Users\WB50674...
10 use "C:\Users\WB506743\D...
11 count
12 br
13 do "C:\Users\WB506743\A...
14 memory
15 do "C:\Users\WB506743\A...
16 do "C:\Users\WB506743\A... 601
17 use "C:\Users\WB506743\D...
18 do "C:\Users\WB506743\A...
19 use "C:\Users\WB506743\D...
20 use "C:\Users\WB506743\D...
21 tab treatment_hh
22 regress treatment_hh harve...
23 tab treatment_hh
24 tab treatment_hh
25 regress treatment_hh harve...
26 tab treatment_hh
27 regress treatment_hh harve...
28 use "C:\Users\WB506743\D...
29 do "C:\Users\WB506743\A...
Review window
. global projectfolder "C:\Users\wb506743\Dropbox\FC Training\June 2018"
. }
. *
. * Project folder globals
. * ---------------------
. global dataWorkFolder "$projectfolder\data"
. *---
. * 2.1 Lab Task 1
. *---
. *Open the data set
. use "$dataWorkFolder\Rwanda_LWH_OI_Panel", clear
. *Open the second data set
. use "$dataWorkFolder\Rwanda_LWH_OI_Panel_12", clear
. exit
end of do-file
Result Window
Command
Command window
Variables
Filter variables here
Name Label
hh_code hh_code
wave Dataset Wave (BL, FUP1, FUP2, FUP3)
year Year in which the data was collected
treatment_hh 1 if HH is treated if ANY of its plot is treated
treatment_site Dummy equal to 1 if the hh was in a treatment site
site_code_bl Site at the baseline
site_code
sample_group Sample 1B or 1C
qid_bl Baseline HH ID (diff to the hh id we use across waves
no_waves_perhh Number of waves in which hh is
treatment_plot1 1 if Plot1 is treated
treatment_plot2 1 if Plot2 is treated
treatment_plot3 1 if Plot3 is treated
treatment_plot4 1 if Plot4 is treated
treatment_plot5 1 if Plot5 is treated
treatment_plot6 1 if Plot6 is treated
treatment_t1 Interaction btw the variable t1 (post intervention tim
treatment_t2 Interaction btw the variable t2 (post intervention tim
Variable window
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Filename Rwanda_LWH_OI_Panel_12.dta
Label
Notes
Variables 4,902
Observations 2,867
Variable and data properties window
C:\Users\WB506743\Dropbox\FC Training\June 2018\Session Materials\Stata Track 1\Lab 2 - Coding for Reproducible Research\last year material\Lab 1 - Track 1 - Stata Basics\Solutions
CAP NUM OVR

Review window

- In the last task we asked you to copy and paste from the result window. That works, but the best practice way to do it is using the review window.

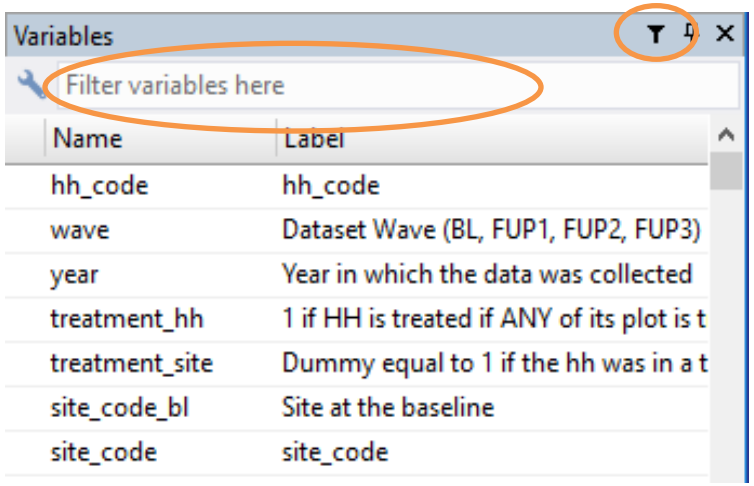
- Double click on a command you want to use again and it will appear in your command window
  - You can also click in command window and select the commands in the result window by using PageUp/PageDown buttons
  - fn+ArrowUp/ fn+ArrowDown on Mac

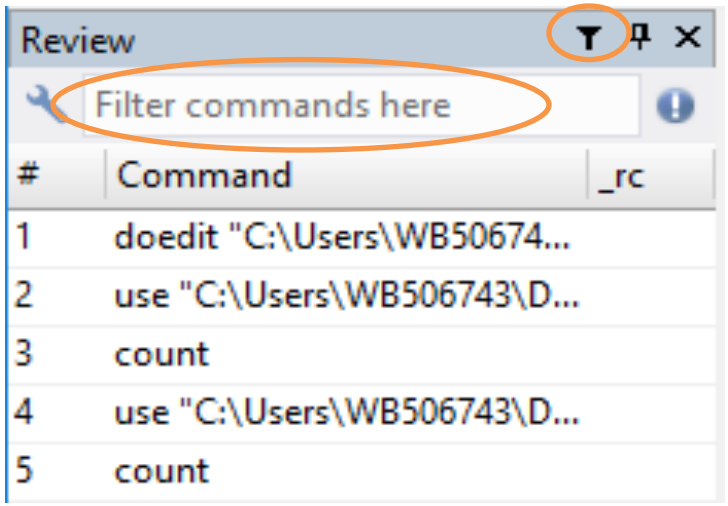
- If a command is red in the review window, it means it hit an error and did not finish

Filter variable and review window

- Both the variable window and the review will soon be very crowded. You can then search both of them

- If you do not see the search bar, click the little funnel symbol





An introduction to Stata:

Exploring a data set opened for the first time

Exploring a new dataset

- *browse*: see all data in spreadsheet format
- *describe*: list of all variables in memory
  - Total number of variables & observations (size of matrix)
  - Variable name, type, format, value label name, variable label
- *summarize*: Basic statistics for numeric variables
  - N, mean, standard deviation, min, max
- tabulate : frequencies

Exploring a new dataset - more

- *codebook*: for each variable, displays:
  - Type (more detail than describe)
  - Number of unique values
  - Number of missing values
  - Range and units
  - Examples of values (strings); tabulations (categorical); or mean, sd and percentiles (continuous)
  - Warnings if embedded blanks (may or may not be ok)
- *labelbook:* for each stored value label, displays:
  - Label definitions
  - what variables labels are applied to
- *list*: lists all variables and observations.
  - Can qualify: *"list if inc_01<500000", "list in 1/10"*
- *summarize, detail*: percentiles, variance, skewness, kurtosis

Types of variables

- In Stata, each variable (column) has to be either:
  - string (text) – values are red when browsing.
  - numeric (number) – values are black or blue when browsing.
- Numbers can be stored as text, but text cannot be stored as number.
  - However, not possible to do computations on numbers stored as text
- Categorical variables should be stored as numeric variables and have labels.

Lab Task 2

- Open *endline_data_raw.dta again.*

- Which variables are string and which are numeric?
  - browse – see the different colors in the columns.
  - *describe* – see the column *storage type.*
  - *summarize* – why is there no mean for the variable "**lwh_group**"?

- Learn more about the variable '**numplots**'. What values does it take on? What is minimum? Maximum? Mean? How many unique values?
  - *tabulate numplots.*
  - *summarize numplots OR summarize numplots, detail.*
  - *codebook numplots.*

An introduction to Stata:

Generating new variables and labeling them

Create new variables

You create new variables using the command *generate*

- Calculate the household income by household size which is the per capita income of all beneficiaries in the sample.
- Generate a dummy variable that is 1 if per capita of income is more than 500,000 RWF.

```
*Generate a variable that is HH total income by size
generate inc_per_hh_member = inc_t / pl_hhsize

*Label variables clearly
label variable inc_per_hh_member "HH total income by HH size"

*Summarize
summarize inc_per_hh_member

*Generate a dummy that 1 if per capita income is
*is greater than than 500,000 francs.
generate high_income = 0
replace  high_income = 1 if inc_per_hh_member > 500000

*Tabulate the result
tabulate high_income
```

Label variables you create

- You will thank yourself if you carefully document what you do. You will not remember everything a month later.

- Label all variables you create, so that future you and others understand.

```
*Label variables clearly
label variable inc_per_hh_member "HH total income by HH size"
```

Value labels

- For all categorical variables, you should also create value labels, to indicate what each category stand for.
- In the example below we apply a label explaining the dummy we just created. This variable will now be very clear to anyone using this data set.

```
*Create the label
label define high_income_lab 1 "Per Capita Income >500K" 0 "Per Capita Income <500K"

*Appy the label to the variable
label value high_income high_income_lab

*Tabulate the variable again, then
*browse it together with HH income by HH size
tabulate high_income
browse   high_income inc_per_hh_member
```

Lab Task 3

1. The variable ***inc_01*** indicates the household income from on-farm enterprise in Rwanda Francs. Create a new variable that depicts it in US dollars. About 0.0012 USD exchanges to 1 Rwanda franc. Call the new variable *inc_01_USD.*

```
*Create a new variable that exchanges HH farm enterprise income to USD
generate inc_01_USD = inc_01 * 0.0012
```

2. Use the same code and create the variable *inc_01_EUR0.* (1 RWF= .0010 Euro.)
3. Label the variable *inc_01_USD* like this:

```
*Label variables maize_tons_annual
label variable inc_01_USD   "On-Farm Enterprise Income (USD)"
```

4. Create a label for *inc_01_EURO* as well.
5. Create a dummy variable for high on-farm enterprise income in the same way *high_income* was created. Use the cut of 100 Euros. Remember to label it!

An introduction to Stata:

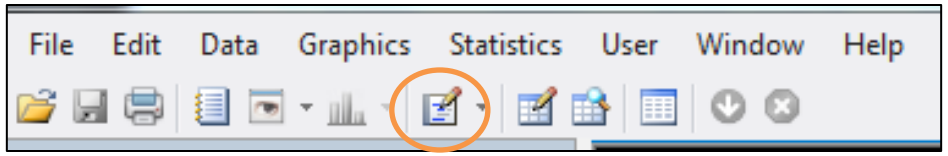
How to share your work with your team

You are asked to share your work

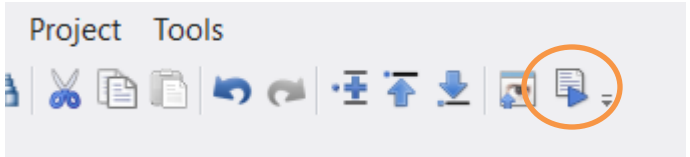
- How would you share the work you have done so far?
- Send only the data set? That would be like Excel and only share the latest version of the data.
- In research we need to share more than the latest version of the data. We need to show what we did.
- Yes, you guessed it, it is time to introduce the do-file.

Do-files

- Open up a new do-file. Window -> Do-file Editor -> New Do-file Editor. Or click the shortcut highlighted below:



- Treat the do-file similarly to how you treat the command window. But instead of copying and running one line of code at the time, a do-file lets you do that with any number of lines of code.
- Running the code in your do-file using menus: Tools -> Execute (Do). Or Ctrl+D (Windows) or this short cut:



Lab Task 4

1. Open a new do-file. Save it! Go to the review window and copy the code where you loaded ***endline_data_raw.dta***. Now run your do-file.

2. Next, use the review window to copy to your do-file all the actions you already did:
   a) Open data *endline_data_raw.dta (done in 1. above).*
   b) Explore the data.
   c) Generate *inc_per_hh_member* and *inc_01_USD.*
   d) Label *inc_per_hh_member* and *inc_01_USD.*
   e) Create and label the *high_income* variable.
   f) Now run your do-file again. You have written instructions how to get to the *endline_data_raw.dta* to the final version.

3. Edit the do-file. Change the cut-off for high HH income from 50,000 to 30,000, and run your do-file again. See how much easier it is to update the instructions instead of making changes to the data!

Comments

- Comments is the green text you have seen in the code examples. Comments is text that Stata will ignore when running your code.

- Comments is what makes the difference between instructions that are easy to follow or impossible to understand.

Different types of comments

1. */* comment */*

   Used for long comments or to explain many lines of code in the following section

2. ** comment*

   Used to explain what happens on the following few rows

3. *// comment*

   Used to explain the same line of code

Other features of Stata

Other features of Stata:

Using macros (globals, locals and scalars)

Macros

- You need to be at least familiar with this topic for the resources you will be introduced to this week.
- This technique is critical as projects grow in size. But even the smallest DIME project absolutely needs this.
- Macros (globals, locals, scalar) save some information (text or number) that you can reference later.
  - Example, we want to access files in the folder multiple times. We can store the folder location in a global and use it multiple times.

Defining macros

```
local numberA 3
local numberB 5

gen restult_var = (`numberA' * `numberB') - `numberA'
```

- This creates a variable that is 12. (3 * 5) – 3 = 12

```
global project GAFSP

gen country_${project}   = "Kenya"
gen donor_${project}     = "World Bank"
```

- This creates two variables.
  - One that is called country_GAFSP and with the value *Kenya* for all observations
  - Another that is called donor_GAFSP with the value *World Bank* for all observations

Setting directories using macros

```
*Load original data
use "C:\Users\wb506743\Dropbox\FC Training\June 2018\data\original\endline_data_raw.dta", clear

    *Work on your data here

*Save final data
save "C:\Users\wb506743\Dropbox\FC Training\June 2018\data\original\endline_data_int.dta", replace
```

- The code above and below is the same to Stata. But the part that is the identical in *use* and *save* is stored in and referenced from a global below

```
*Set folder global
global folder_labl  "$projectfolder\data\original"

*Load original data
use "$folder_labl\endline_data_raw.dta", clear

    *Work on your data here

save "$folder_labl\endline_data_int.dta", replace
```

Lab Task 5

1. Open the do-file you created in the previous task and create a global to the folder where you have been saving your work. Call the global *folder_Lab1*
2. To check what is stored in your global, type this in the command window and hit enter:
   - *display "$folder_Lab1“*
3. Update the *use* and *save* commands in your do-file with the global you created. Like this:

```
*Load original data
use "$dataWorkFolder\endline_data_raw.dta", clear
```

4. Now, run your do-file again and test that it works

Other features of Stata:

White space

White Space

- Stata does not distinguish between one empty space and many empty spaces, or one line break or many line breaks

- It makes a big difference to the human eye and we would never share a Word document, an Excel sheet or a PowerPoint presentation without thinking about white space – although we call it formatting

Vertical lines

```
gen NoPlotDataBL = 0
replace NoPlotDataBL = 1 if c_plots_total_area >= .

gen NoHarvValueDataBL = 0
replace NoHarvValueDataBL = 1 if c_harv_value >= .

rename c_gross_yield c1_gross_yield
rename c_net_yield  c1_net_yield
rename c_harv_value c1_harv_value
rename c_total_earnings c1_total_earnings
rename c_input_spend c2_inp_total_spending
rename c_IAAP_harv_value c1_IAAP_harv_value
rename c_plots_total_area  c1_total_plotsize
rename  c1_cropPlotShare_??? c1_cropPlotShare_all_???

tempfile BL_append
save  `BL_append'
```

```
gen     NoPlotDataBL = 0
replace NoPlotDataBL = 1        if c_plots_total_area >= .

gen     NoHarvValueDataBL = 0
replace NoHarvValueDataBL = 1   if c_harv_value >= .

rename  c_gross_yield           c1_gross_yield
rename  c_net_yield             c1_net_yield
rename  c_harv_value            c1_harv_value
rename  c_total_earnings        c1_total_earnings
rename  c_input_spend           c2_inp_total_spending
rename  c_IAAP_harv_value       c1_IAAP_harv_value
rename  c_plots_total_area      c1_total_plotsize

rename  c1_cropPlotShare_???    c1_cropPlotShare_all_???

tempfile BL_append
save    `BL_append'
```

Vertical lines

```
*-create dummy for employed
gen employed = 1
replace employed = 0 if _merge == 2
label var employed "Person exists in employment data"
label define yesno 1 "yes" 0 "no"
label val employed yesno
```

```
*-create dummy for being employed
gen         employed = 1
replace     employed = 0 if _merge == 2
label var   employed "Person exists in employment data"
label def            yesno 1 "yes" 0 "no"
label val   employed yesno
```

Other features of Stata:

Missing values

Missing values

- String variables can be empty, but numeric variables can't be empty. Instead numeric variables have something called "missing values".
  - Missing values are represented in Stata with a dot, like this: .
  - You can also use *.a* or *.b* etc. to *.z* for missing values and you will learn later how these can be used.

- Stata can't use missing values in computations (averages, regressions etc.) so it skips observations with missing values.

- Missing values changes the analysis as observations with missing values are excluded from commands like *summarize* and *regress.*

- Good practice to always check for missing values when tabulating variables.

Lab Task 6

1. Make sure that you have the endline_*data_raw.dta .dta* open.
   - Type summarize inc_* into the command window . Look at the column *obs* in the output. What do you notice?
   - There are 1,067 observations in the data set, but less than 1,067 observations were included in the output for the income variables. Why is that?

2. *misstable summarize* identifies and reports missing values in all the variables in your data set. Use the *misstable summarize* command on the income variables. It reports exactly 105 missing observations for all of the household income variables. That explains all observations missing in the *summarize* output in number 1 above.

4. Another way to investigate the missing observations is to tabulate variable *education_1.*
   - *Test first: tabulate inc_zero.*
   - *Then test: tabulate inc_zero, missing.*

An introduction to Stata:

Saving your data sets

Saving Stata datasets/Lab Task 7

1. Use the drop down menu to save your data under a different name, for example *endline_data_int.dta*. File -> Save as. If you use Save instead of Save as you will not be asked to save under a new name
2. Copy the code generated when using the menu to your do-file. Run again. Did you get an error?
   - Since you already saved the data set under the new name, Stata warns you that there already is a file with that name. We want to overwrite it as each time we make an update to the instructions, we want the final data to reflect that. We therefore need to add a comma and the word *replace*. Like this:

```
*Save final data
save "$dataWorkFolder\endline_data_int.dta", replace
```

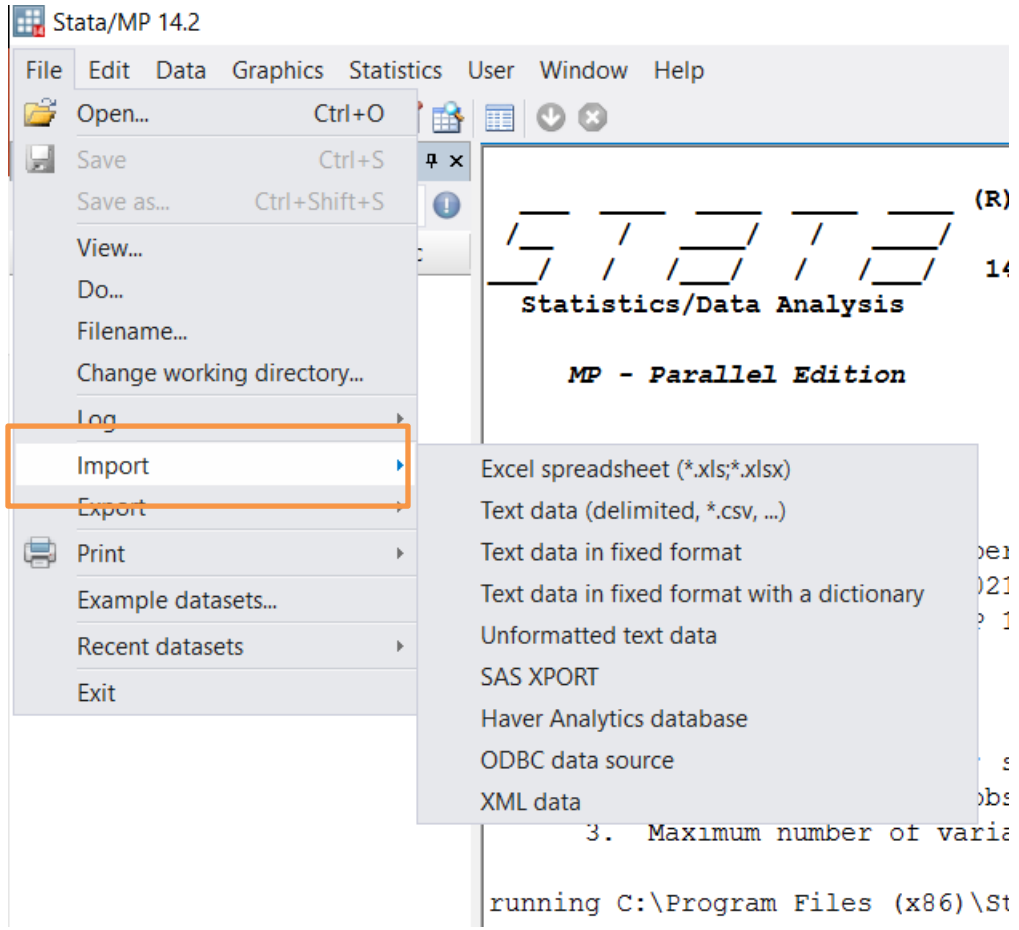
3. Run your code again. Now you both have a final version of your data set and the instructions on how to get there. When your supervisor wants to see your work, share both of these!

Other features of Stata:

Open non-Stata data sets

Importing data from .xls/.xslx or .csv

- Always start by using the drop down menus if you do not know the command!
- Select the format of your file:
  - Excel (.xls,.xlsx)
  - plain text delimited (.txt)
  - comma separated (.csv)
- Select the appropriate choice in the menu and follow the instructions.
- Copy the code Stata Generates for you and put it in your do-file.



Lab Task 8

- Use the drop down menus to import the *village_codes.xls* data to Stata. After you select Excel Spreadsheet you get a new window with all *import excel* options. For now, only use browse to select the file and then click OK.
  - Use *describe*, *tabulate*, *browse* or any other tools you have learnt to explore the data
  - Does everything look ok?

- In excel the first row can be used as variable names, but in Stata the first row is always data. Use the drop down menus to open up the window with the *import excel* options again. Is there any option you see that you think will solve this?
  - *Use the option you think will solve this. Then explore the data again.*
  - *See that the columns that stored text in Excel are imported as strings, and the columns that stored numbers in Excel are imported as numeric*

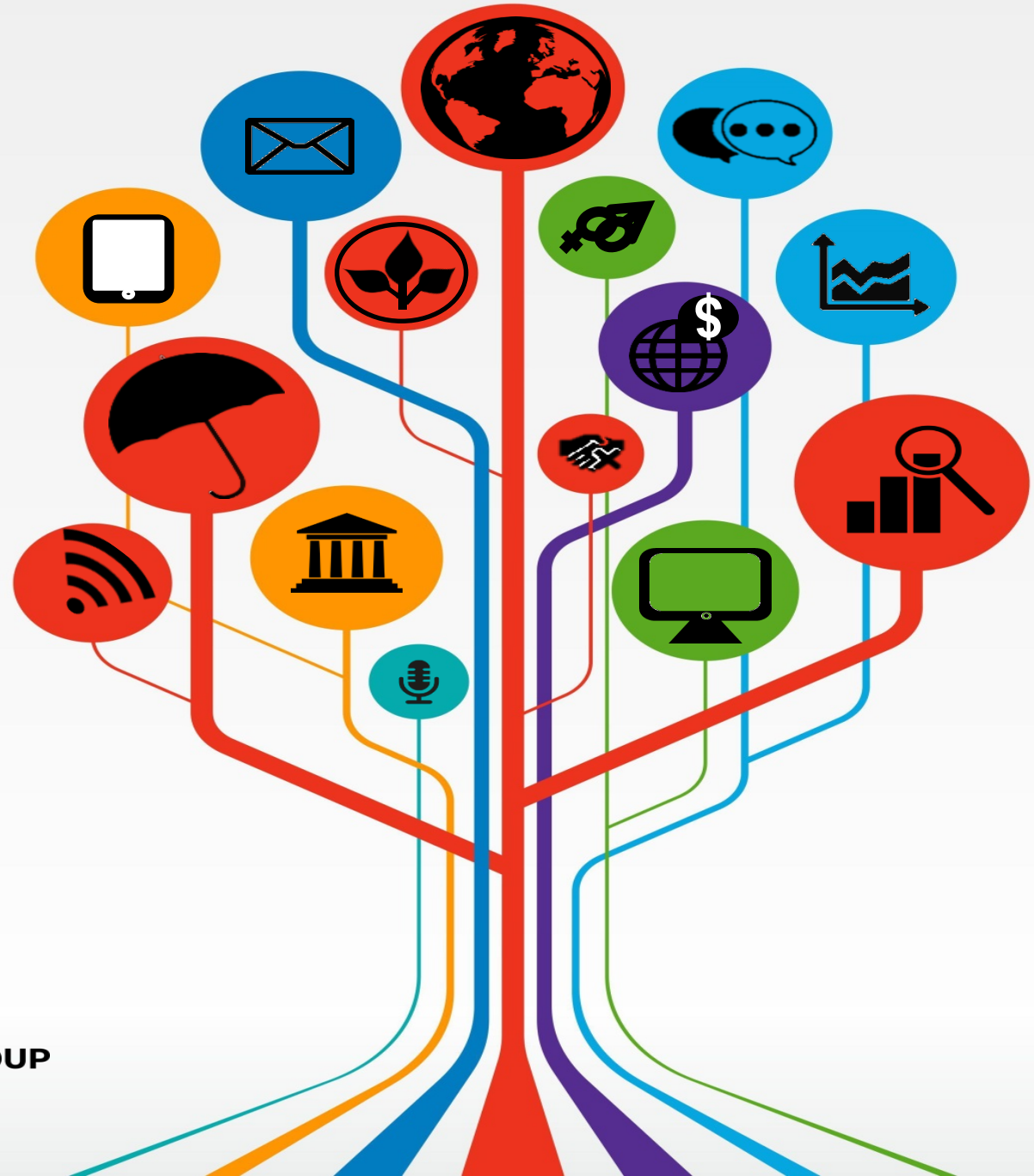
- In the result window or in the review window, see that the two lines of code generated are similar, but have the difference that the second time the word *firstrow* is used.

- It is a very good practice to use the drop down menus and the command window to experiment with your code until you are happy with it. When you are happy with the code Stata generates to import *village_code.xls, c*opy that line of code to your do-file and use the global you created in Task 5 to shorten the file path

Thank you!

Michael Orevba & Kristoffer Bjärkefur

19 June 2018



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP