Surviving Stata Workshop

infoguides.gmu.edu/software/stata

dsc.gmu.edu/software

Please open Stata

Statistical Software

Easy to start, limited capability

Best for those with basic needs or who analyze experiments

Easy flexible syntax, extensible

Best for academics, especially in economics & public policy

Hard to learn, highly capable

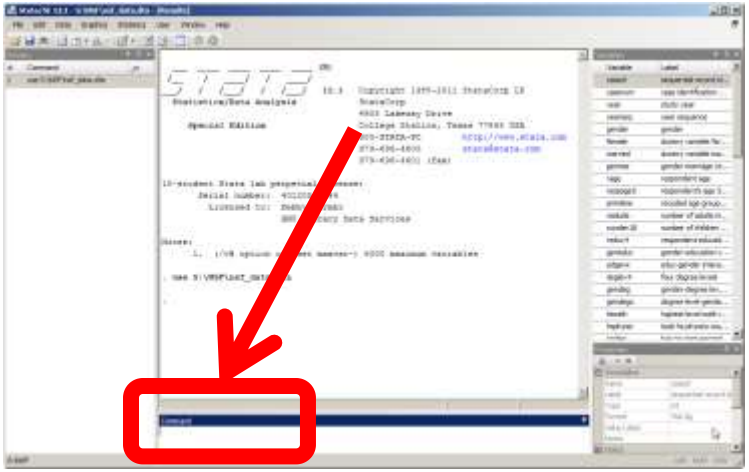
Best for managing huge and/or complex data (govt, financial)

Improving, highly extensible

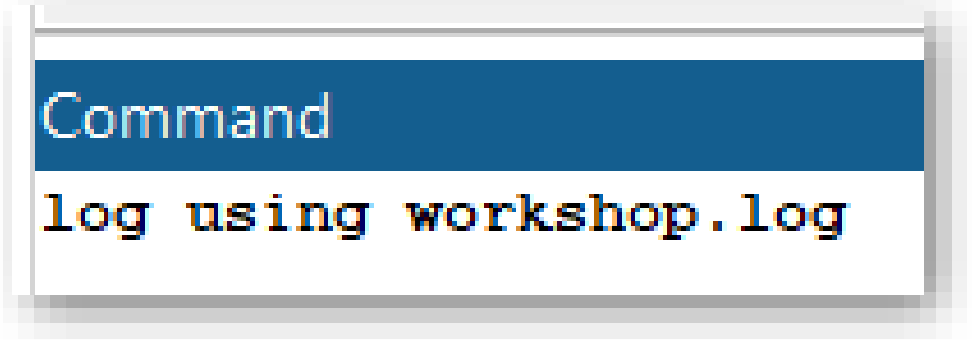
Best for those who program and are comfortable with data

Submitting Stata Commands

1. Click in the
Command Window



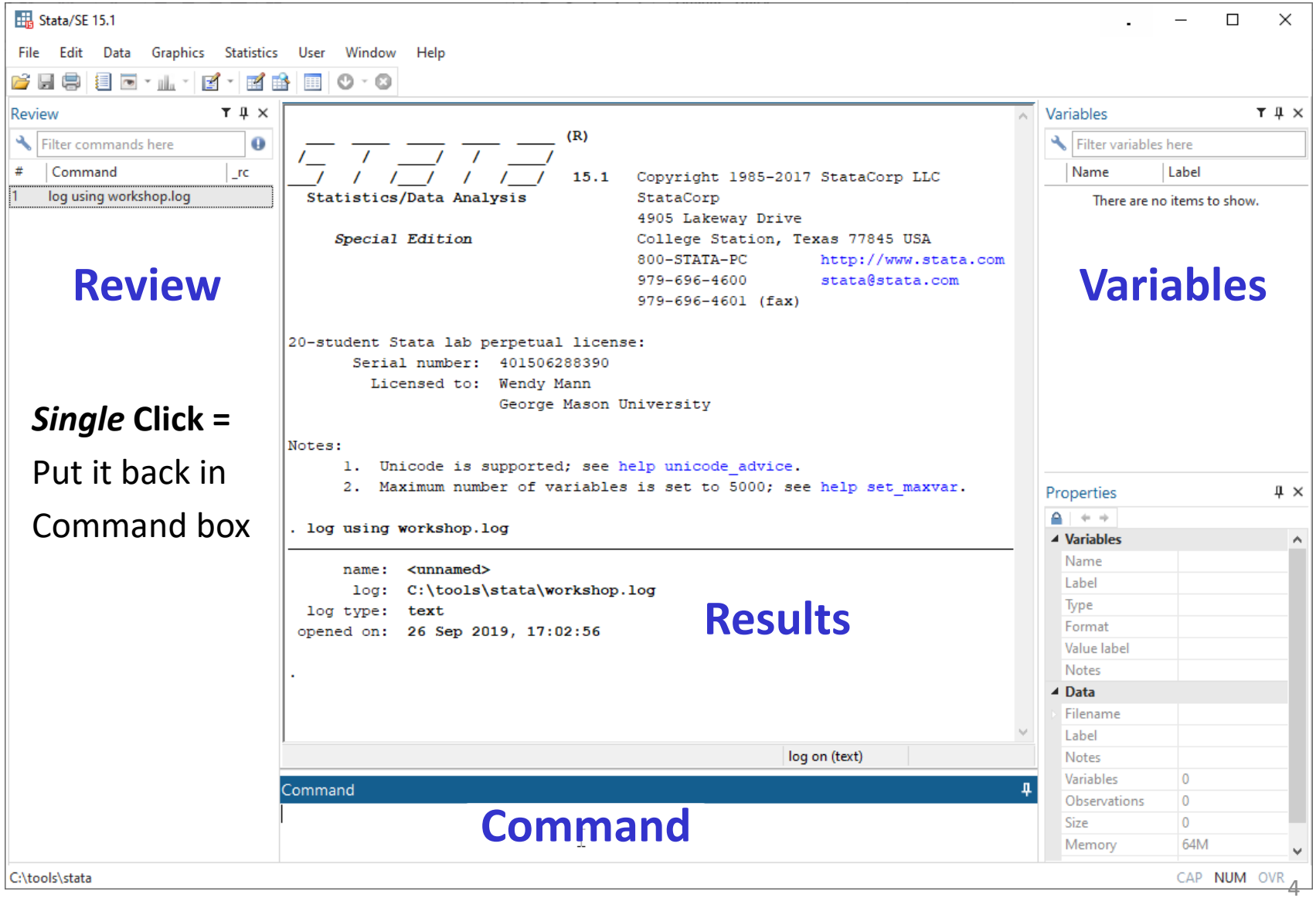
2. Type:
`log using workshop.log`



3. Press **Enter**

```
. log using workshop.log
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\tools\stata\workshop.log
  log type:  text
```

Stata Window



Opening Files

Those participating in the workshop will be provided the data file.

The following instructions enable others to get and open the file.

Getting the Data

1. Register for an account at

   https://www.pewsocialtrends.org/profile/registration/

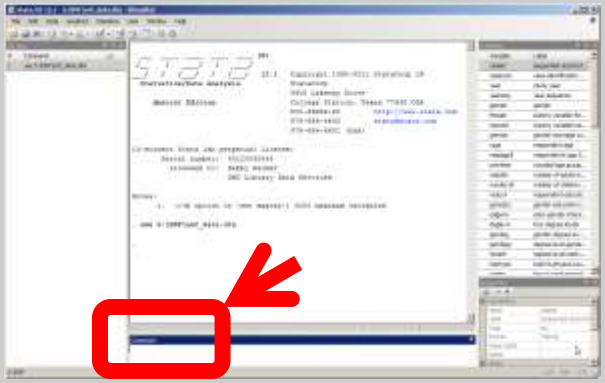
2. Download the datafile from:

   https://www.pewsocialtrends.org/2007/11/09/family-bonds-data/

3. Unzip the folder to a location you can find again

4. Open Stata

5. Paste or type the following in the Command box; press Enter



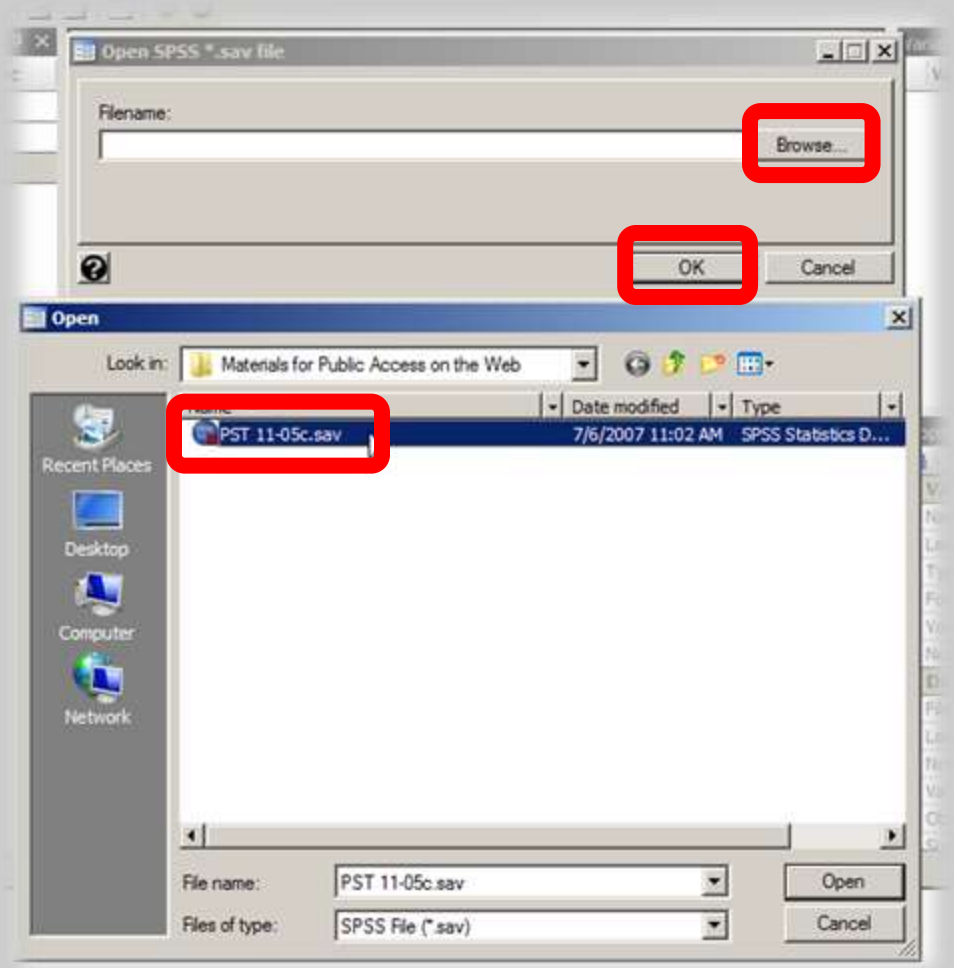
```
net from http://radyakin.org/transfer/usespss/beta
net install usespss
usespss
```

Select the File

Press "Browse…"
in the dialog box

Find and select
the data file (.sav)

Press OK

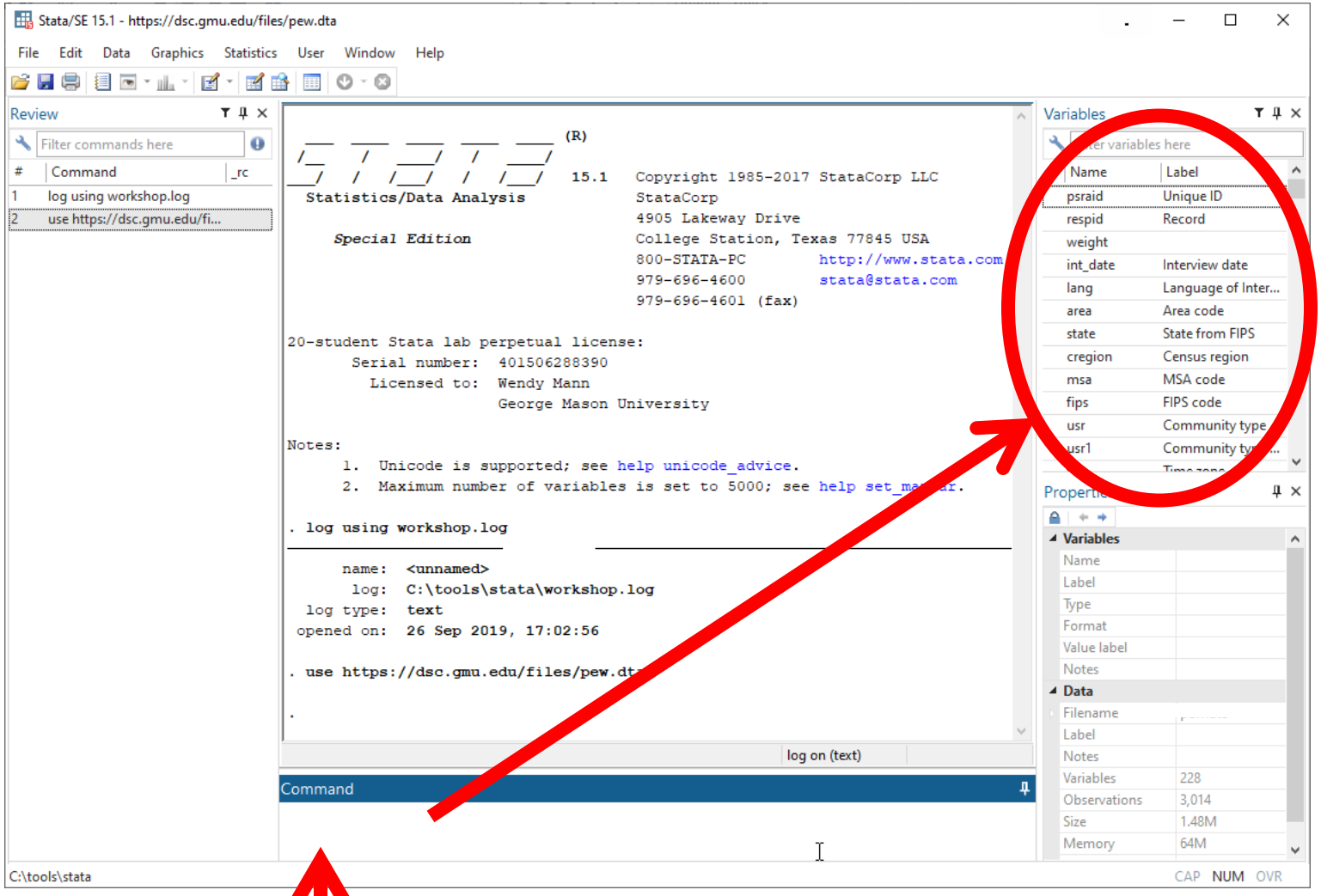


Stata Vocabulary

```
log using workshop.log
```

```
use https://dsc.gmu.edu/files/
```

use https://dsc.gmu.edu/files/pew.dta



Command

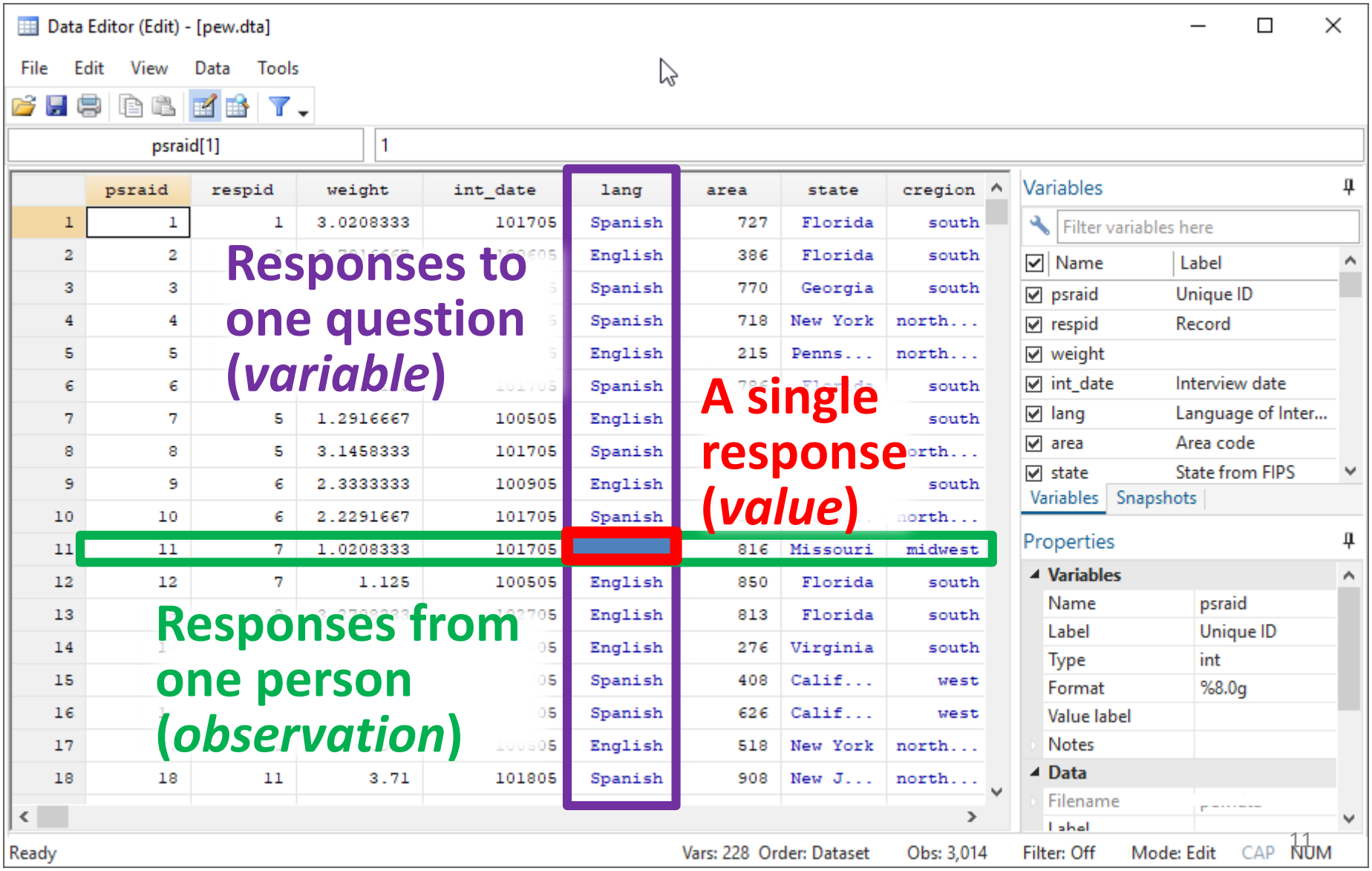
```
use https://dsc.gmu.edu/files
```

Looking at Data

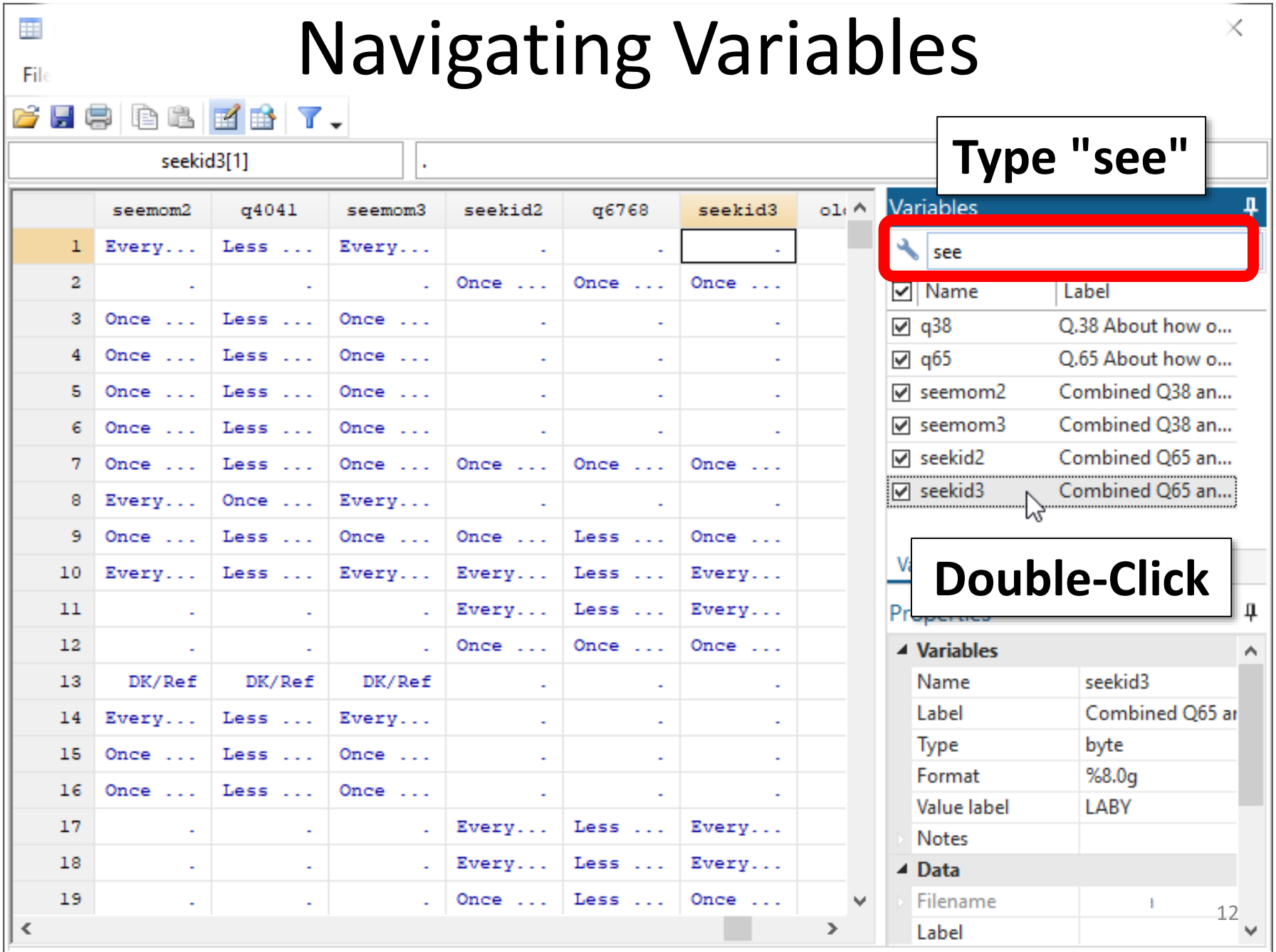
browse

Pew Research Center. (2007). Pew Social Trends [Data file]. Retrieved from http://www.pewsocialtrends.org/category/datasets/?download=5753

browse

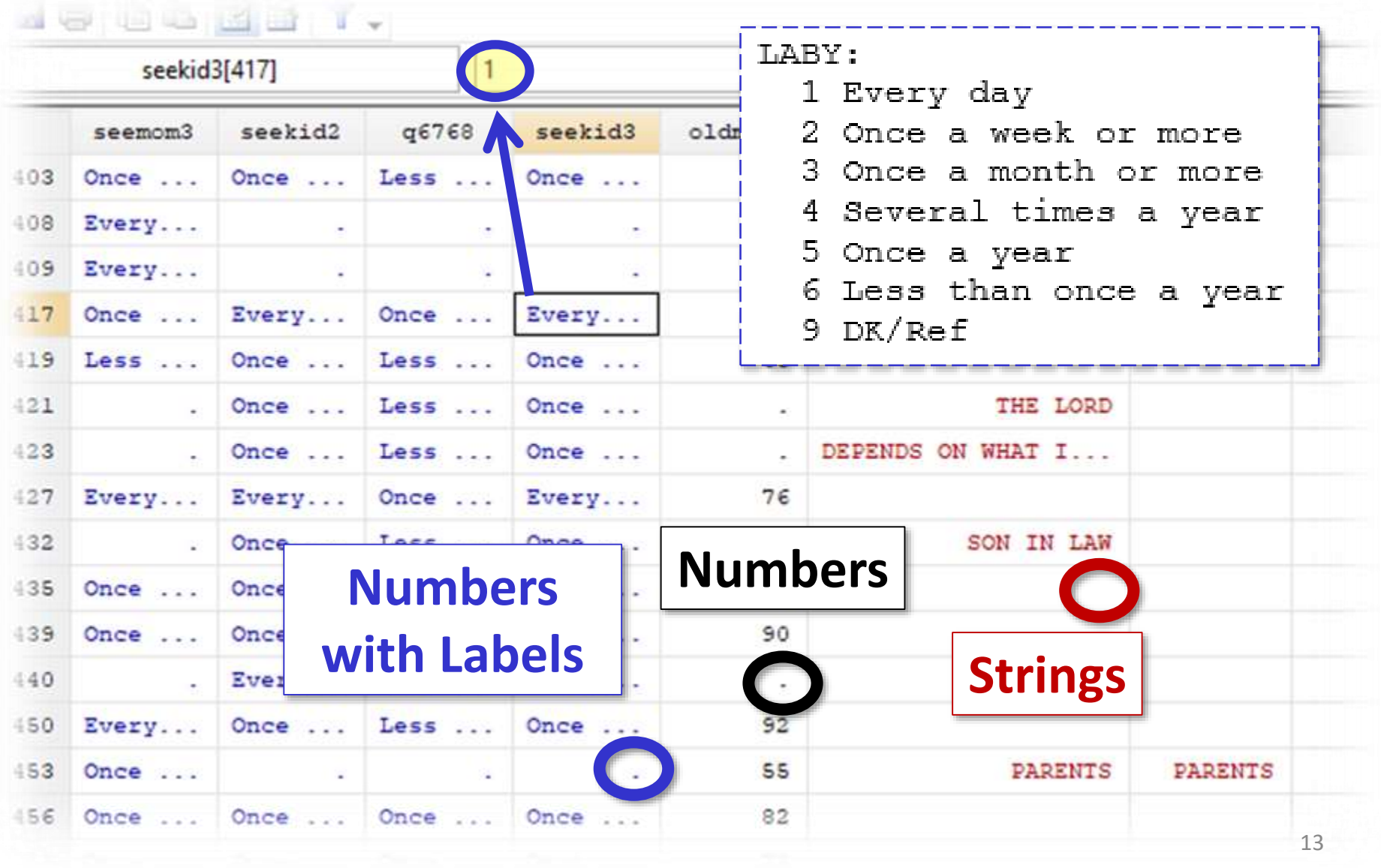

Data Editor (Edit) - [pew.dta]
File Edit View Data Tools
psraid[1]
1
psraid respid weight int_date lang area state cregion
1 1 1 3.0208333 101705 Spanish 727 Florida south
2 2 English 386 Florida south
3 3 Spanish 770 Georgia south
4 4 Spanish 718 New York north...
5 5 English 215 Penns... north...
6 6 Spanish south
7 7 5 1.2916667 100505 English south
8 8 5 3.1458333 101705 Spanish
9 9 6 2.3333333 100905 English south
10 10 6 2.2291667 101705 Spanish
11 11 7 1.0208333 101705 816 Missouri midwest
12 12 7 1.125 100505 English 850 Florida south
13 English 813 Florida south
14 English 276 Virginia south
15 Spanish 408 Calif... west
16 Spanish 626 Calif... west
17 English 518 New York north...
18 18 11 3.71 101805 Spanish 908 New J... north...
Responses to one question (variable)
A single response (value)
Responses from one person (observation)
Variables
Filter variables here
Name Label
psraid Unique ID
respid Record
weight
int_date Interview date
lang Language of Inter...
area Area code
state State from FIPS
Variables Snapshots
Properties
Variables
Name psraid
Label Unique ID
Type int
Format %8.0g
Value label
Notes
Data
Filename
Label
Ready
Vars: 228 Order: Dataset
Obs: 3,014
Filter: Off
Mode: Edit
CAP NUM

Navigating Variables

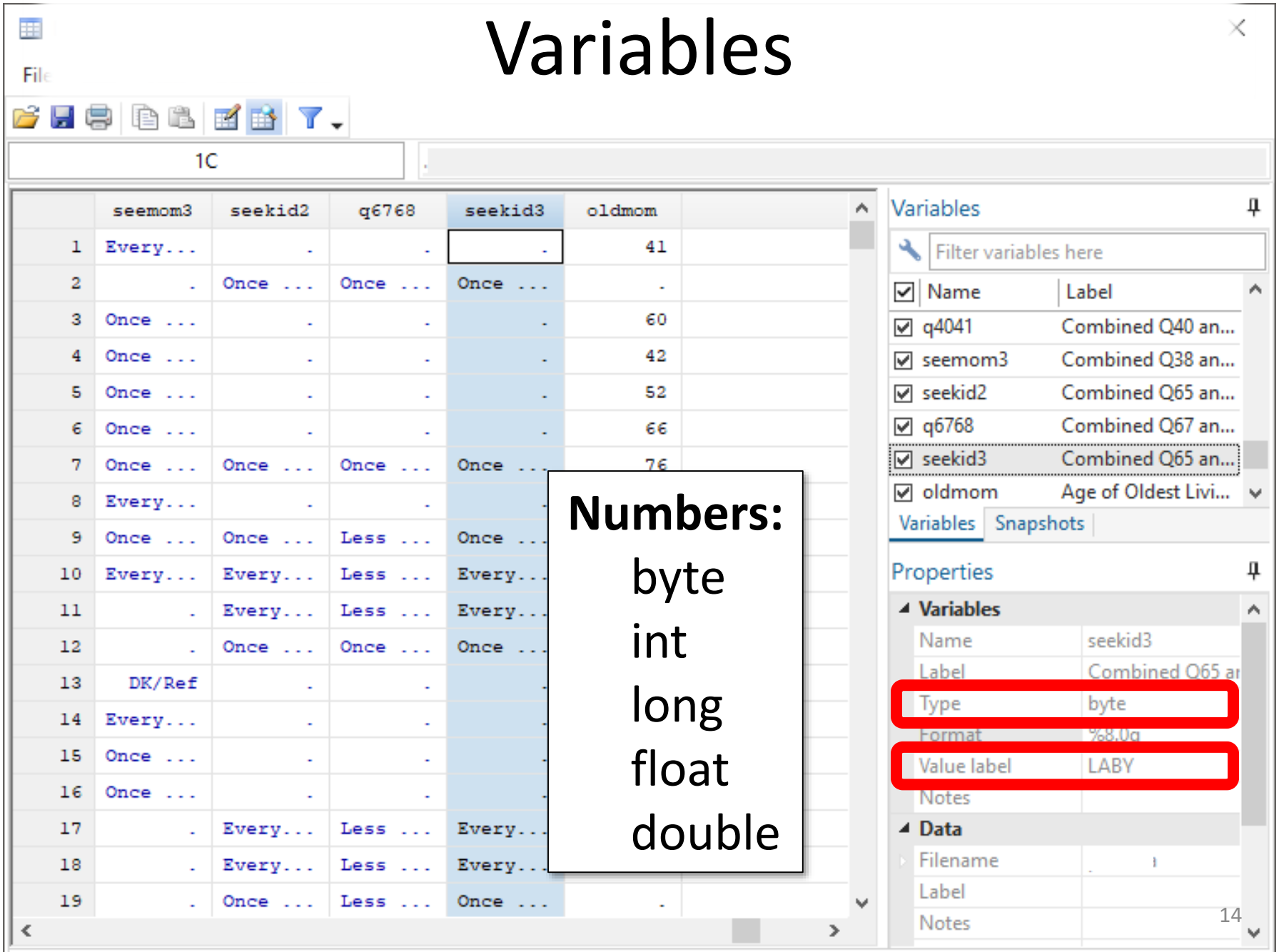


	seemom2	q4041	seemom3	seekid2	q6768	seekid3
1	Every...	Less ...	Every...	.	.	.
2	.	.	.	Once ...	Once ...	Once ...
3	Once ...	Less ...	Once ...	.	.	.
4	Once ...	Less ...	Once ...	.	.	.
5	Once ...	Less ...	Once ...	.	.	.
6	Once ...	Less ...	Once ...	.	.	.
7	Once ...	Less ...	Once ...	Once ...	Once ...	Once ...
8	Every...	Once ...	Every...	.	.	.
9	Once ...	Less ...	Once ...	Once ...	Less ...	Once ...
10	Every...	Less ...	Every...	Every...	Less ...	Every...
11	.	.	.	Every...	Less ...	Every...
12	.	.	.	Once ...	Once ...	Once ...
13	DK/Ref	DK/Ref	DK/Ref	.	.	.
14	Every...	Less ...	Every...	.	.	.
15	Once ...	Less ...	Once ...	.	.	.
16	Once ...	Less ...	Once ...	.	.	.
17	.	.	.	Every...	Less ...	Every...
18	.	.	.	Every...	Less ...	Every...
19	.	.	.	Once ...	Less ...	Once ...

Values


seekid3[417]
1
LABY:
1 Every day
2 Once a week or more
3 Once a month or more
4 Several times a year
5 Once a year
6 Less than once a year
9 DK/Ref
seemom3 seekid2 q6768 seekid3 oldr
403 Once ... Once ... Less ... Once ...
408 Every... . . .
409 Every... . . .
417 Once ... Every... Once ... Every...
419 Less ... Once ... Less ... Once ...
421 . Once ... Less ... Once ... . THE LORD
423 . Once ... Less ... Once ... . DEPENDS ON WHAT I...
427 Every... Every... Once ... Every... 76
432 . Once ... . SON IN LAW
435 Once ... Once ...
439 Once ... Once ... 90
440 . Ever ... .
450 Every... Once ... Less ... Once ... 92
453 Once ... . . . 55 PARENTS PARENTS
456 Once ... Once ... Once ... Once ... 82
Numbers with Labels
Numbers
Strings

Variables



Numbers:

byte

int

long

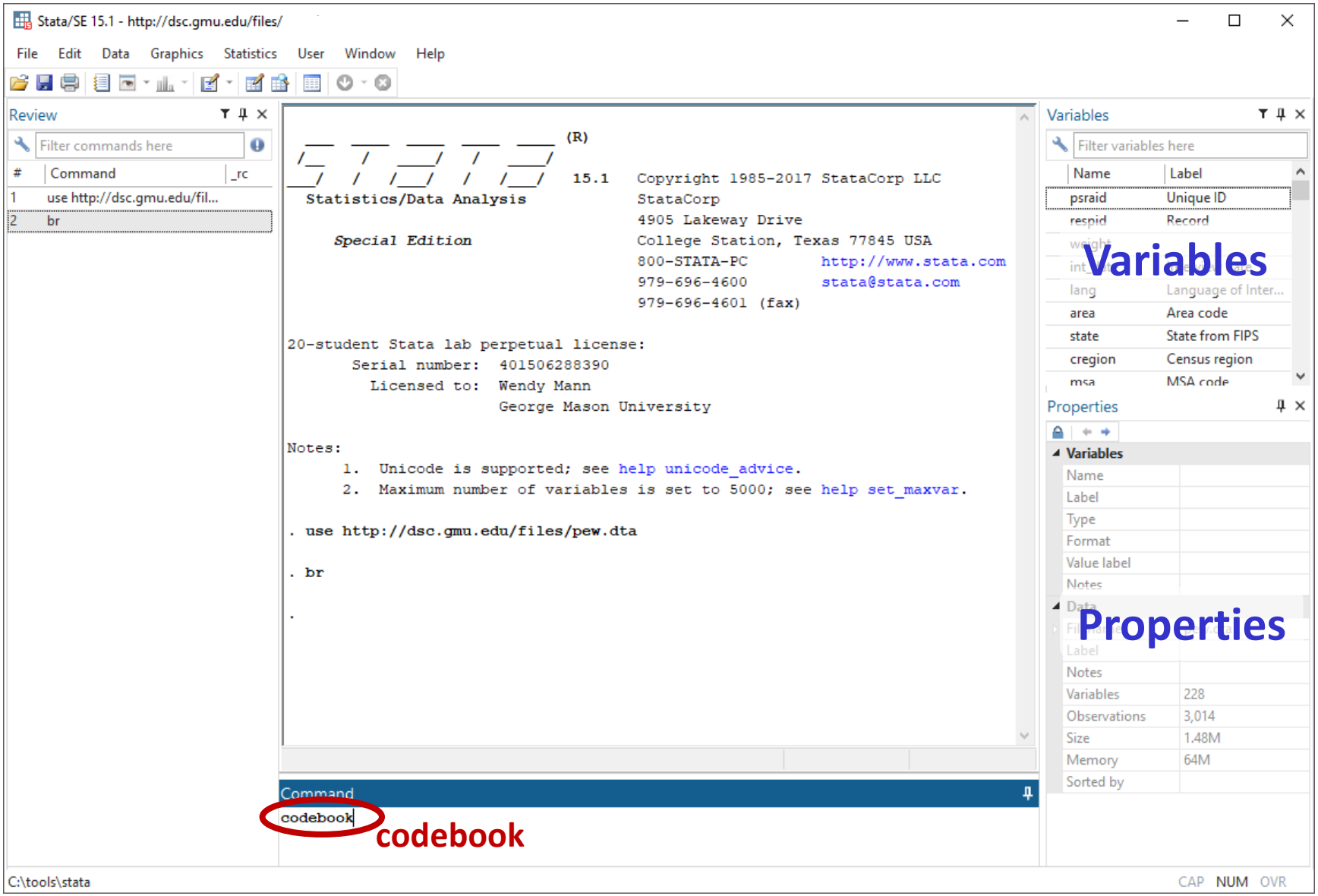
float

double

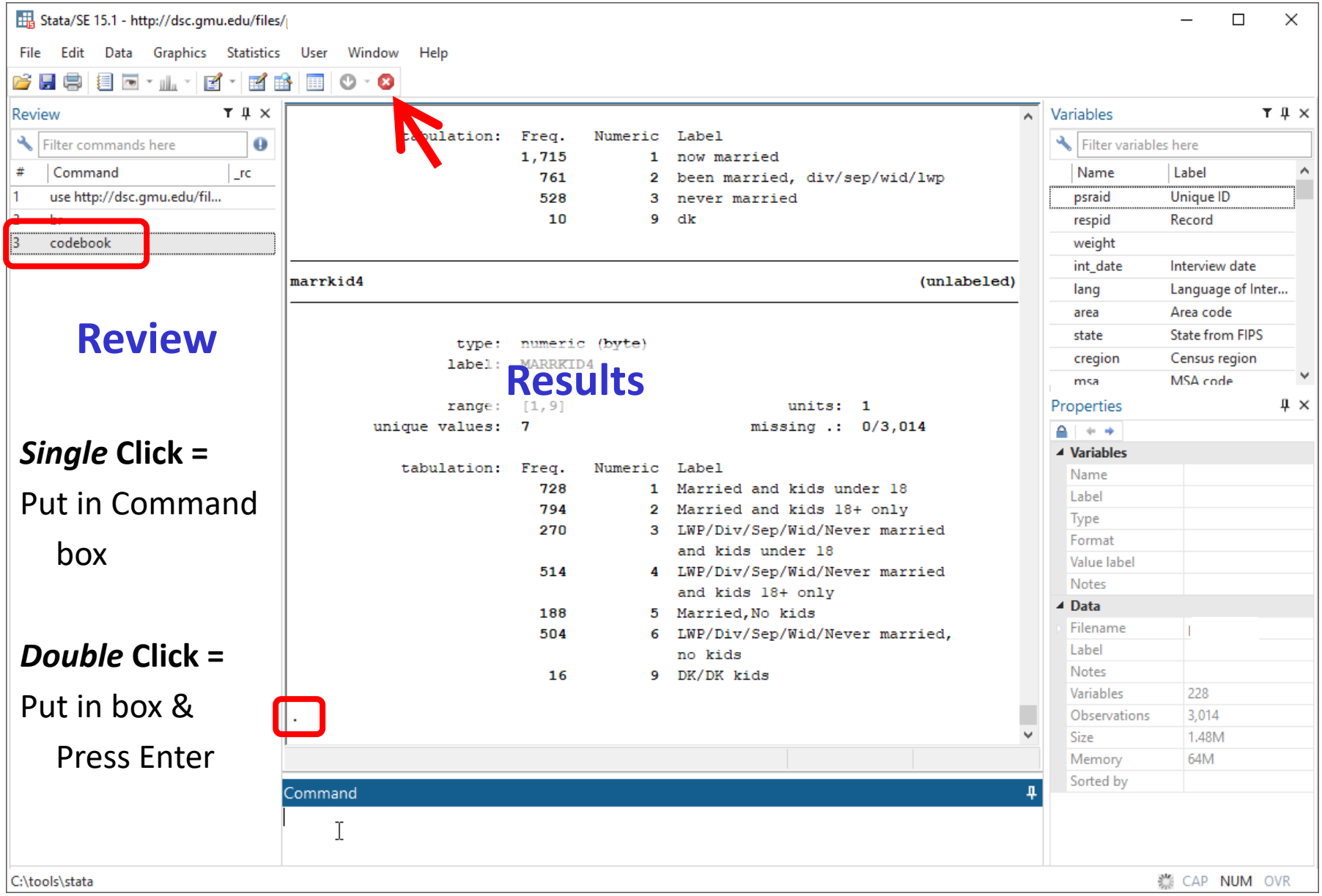
Interface & Windows

Close the browse window

Main Window


Stata/SE 15.1 - http://dsc.gmu.edu/files/
File Edit Data Graphics Statistics User Window Help
Review
Filter commands here
Command _rc
1 use http://dsc.gmu.edu/fil...
2 br
(R)
Statistics/Data Analysis
15.1
Special Edition
Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
20-student Stata lab perpetual license:
Serial number: 401506288390
Licensed to: Wendy Mann
George Mason University
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
. use http://dsc.gmu.edu/files/pew.dta
. br
.
Command
codebook
codebook
Variables
Filter variables here
Name Label
psraid Unique ID
respid Record
area Area code
state State from FIPS
cregion Census region
Variables
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Label
Notes
Variables 228
Observations 3,014
Size 1.48M
Memory 64M
Sorted by
Variables
Properties
C:\tools\stata
CAP NUM OVR

Results



Review

***Single* Click =**

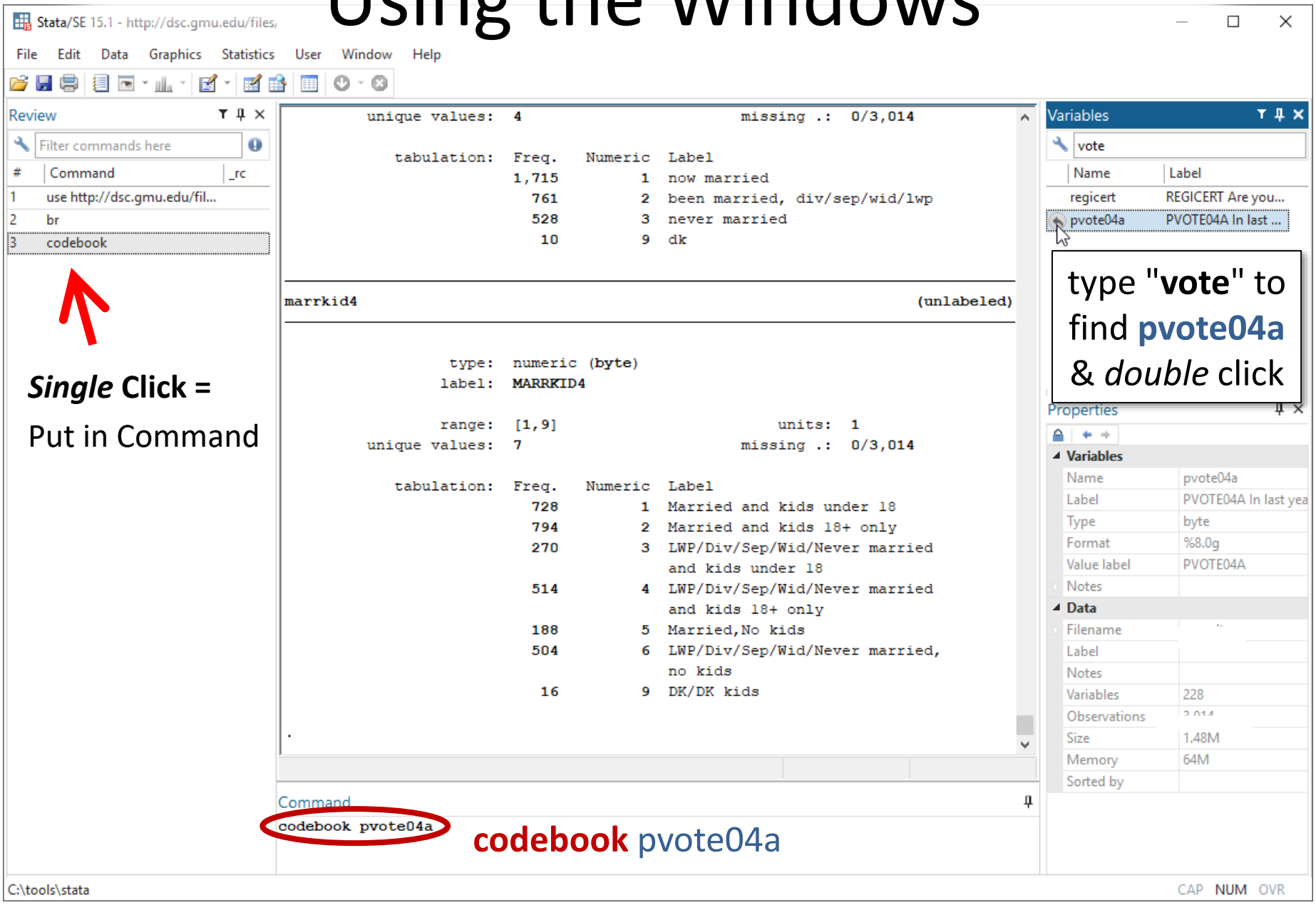
Put in Command box

***Double* Click =**

Put in box & Press Enter

Results

Using the Windows



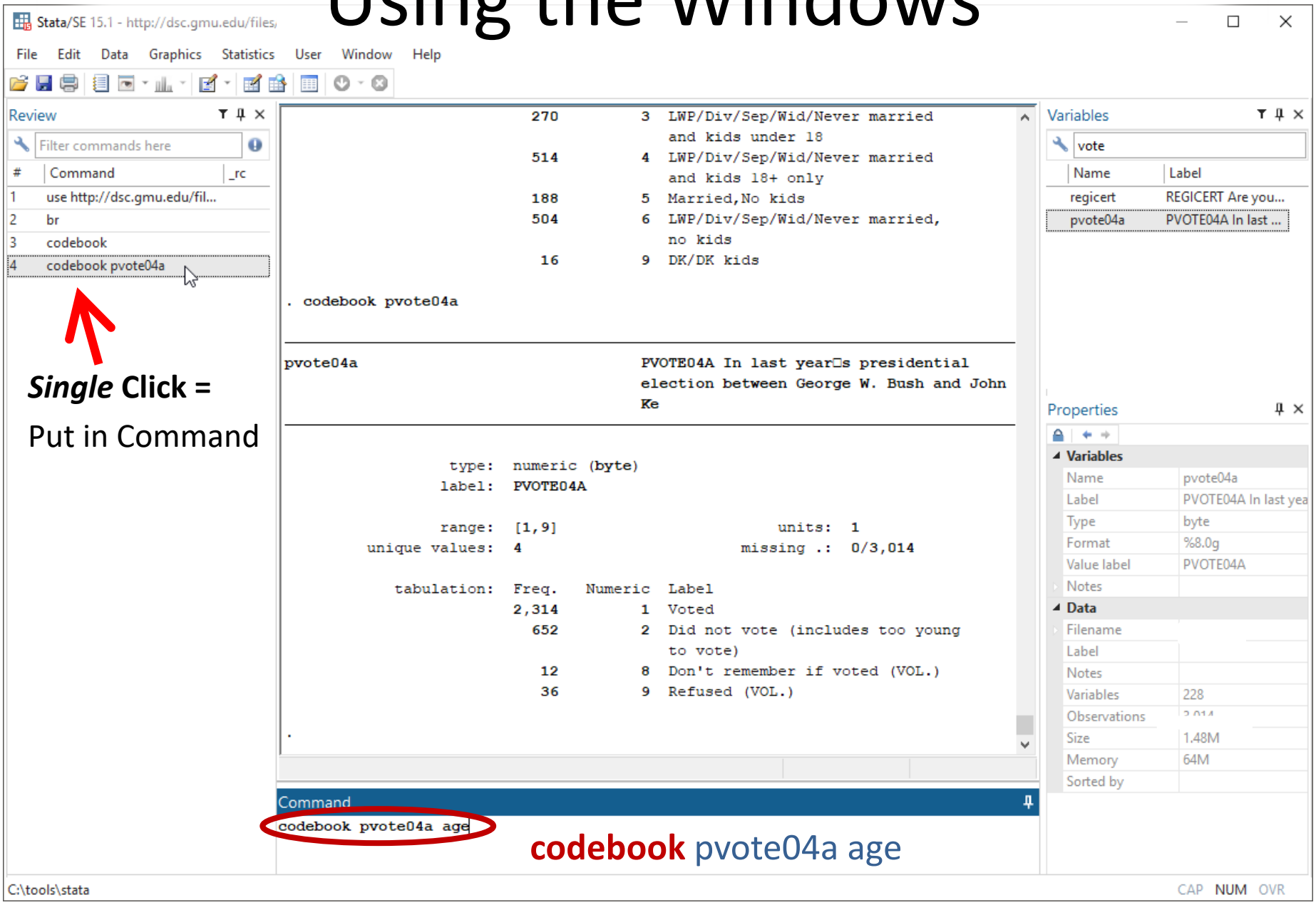
***Single* Click =**

Put in Command

type "**vote**" to find **pvote04a** & *double* click

codebook pvote04a

Using the Windows



Interpreting the Codebook

#1: Black, Red, or Blue?

Storage **Type**

Label Assigned

#2: Category or Measure?

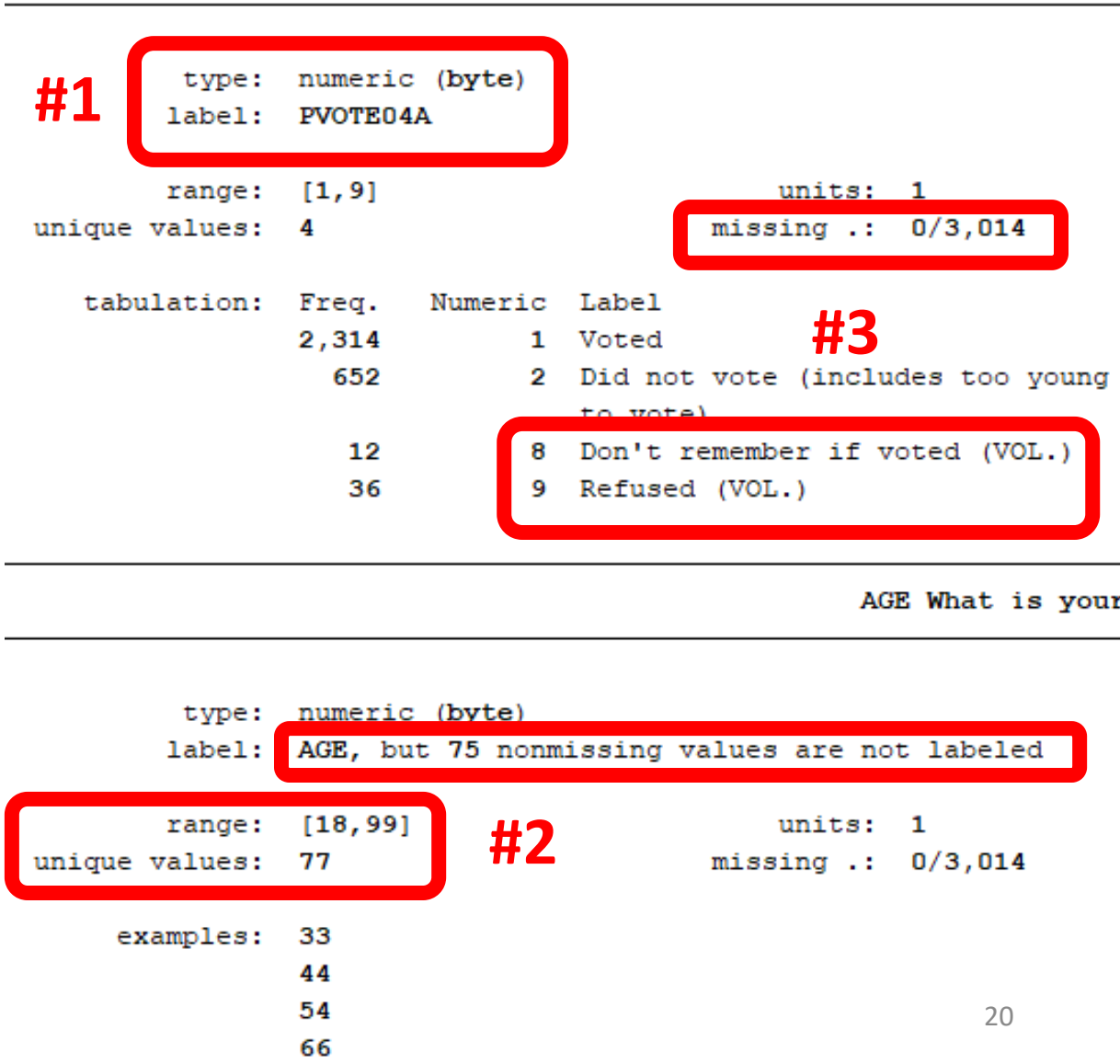
of **Unique Values**

of Values **Labeled**

#3: Missing Values?

True **Missing**

Researcher



Other Variable Types

Black:
Numeric,
no label

Red:
String

```
. codebook fips usr

-------------------------------------------------------------------------------
fips                                                                    FIPS
-------------------------------------------------------------------------------

                  type:  numeric (long)

                 range:  [1001,56041]                 units:  1
         unique values:  1164                     missing .:  0/3014

                  mean:   28406.9
             std. dev:   15725.7

           percentiles:        10%       25%       50%       75%       90%
                              6059     13139     29095     42003     48439

-------------------------------------------------------------------------------
usr                                                     Community type - a
-------------------------------------------------------------------------------

                  type:  string (str1)

         unique values:  3                        missing "":  0/3014

            tabulation:  Freq.  Value
                          640  "R"
                         1509  "S"
```

Syntax Structure

Syntax

command varlist qualifiers, options

histogram age**,** normal width**(**1**)**

tabulate age **if** **(**age **<** 99**)**

label list AGE

generate yrborn **=** 2005 - age

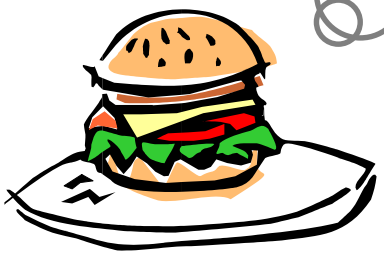
histogram yrborn **if** **(**age **<** 99**),** normal

Structure

order burger if (price < 3), plain mustard

order burger, meat(2) cheese(American)

order burger if (place != "Carls Jr"), nomayo



Abbreviations

histogram age, **norm**al **w**idth(1)

hist age, **norm** **w**(1)

tabulate yrborn

tab yrborn

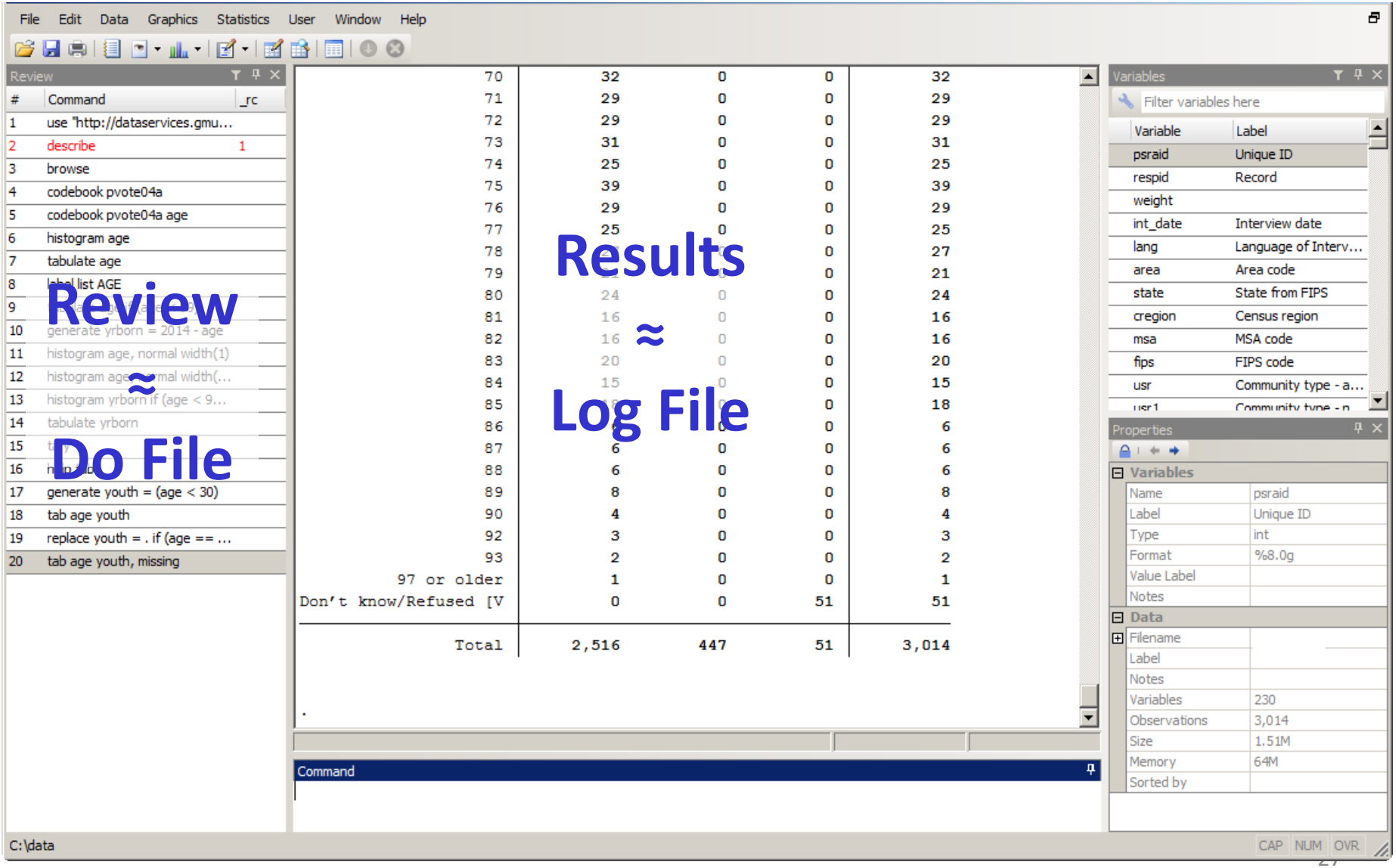
tab yrb

ta y

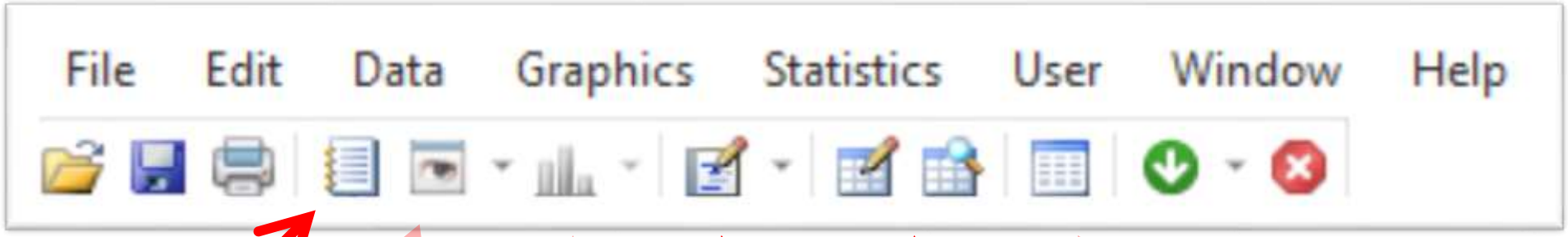
Wildcards
? = One
***** = Multiple

Keeping Track

Do Files & Logs



Menu Bar



Logs
begin/ close

Viewer
help & smcl logs

Graphs
use names to have multiple windows

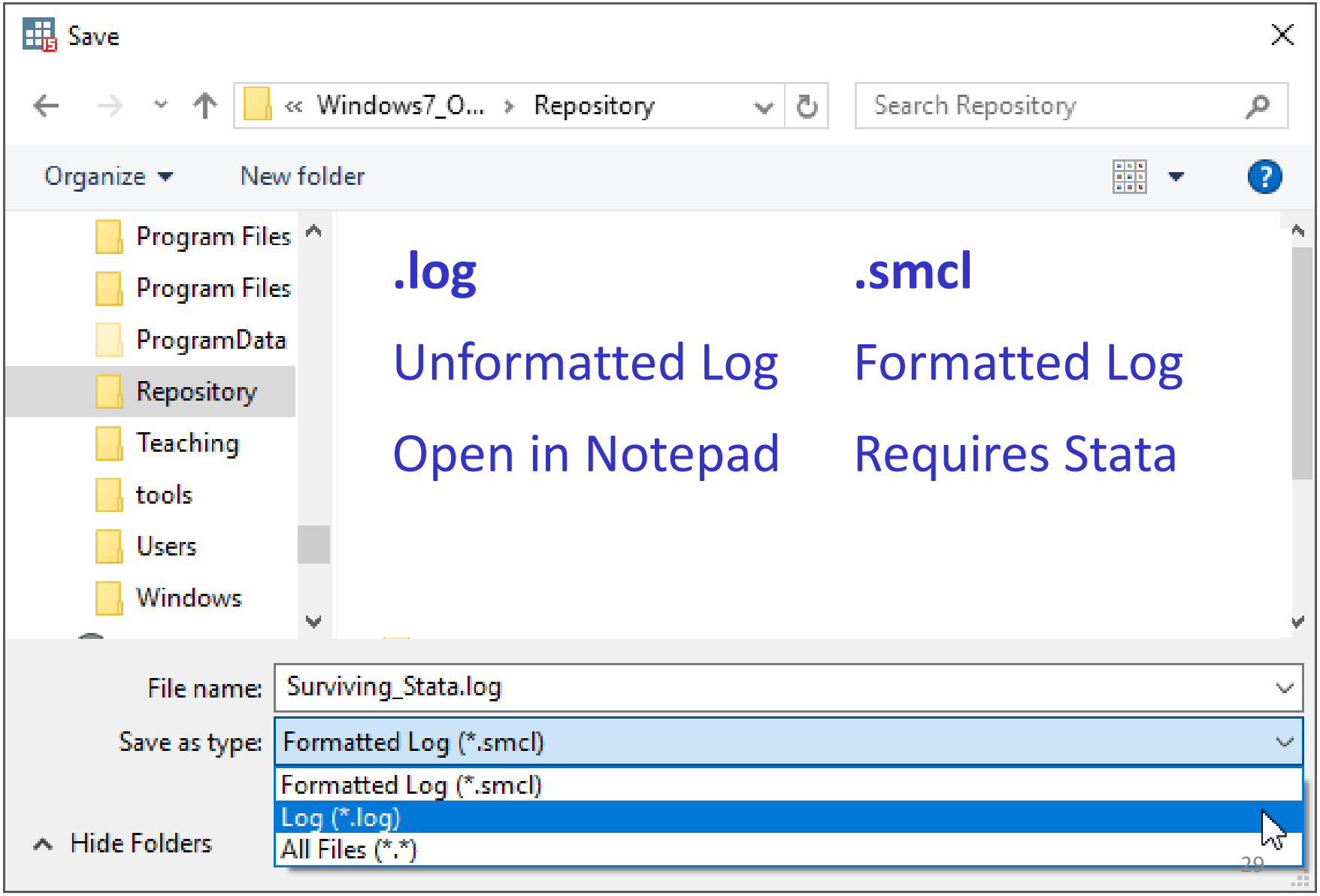
Do Files

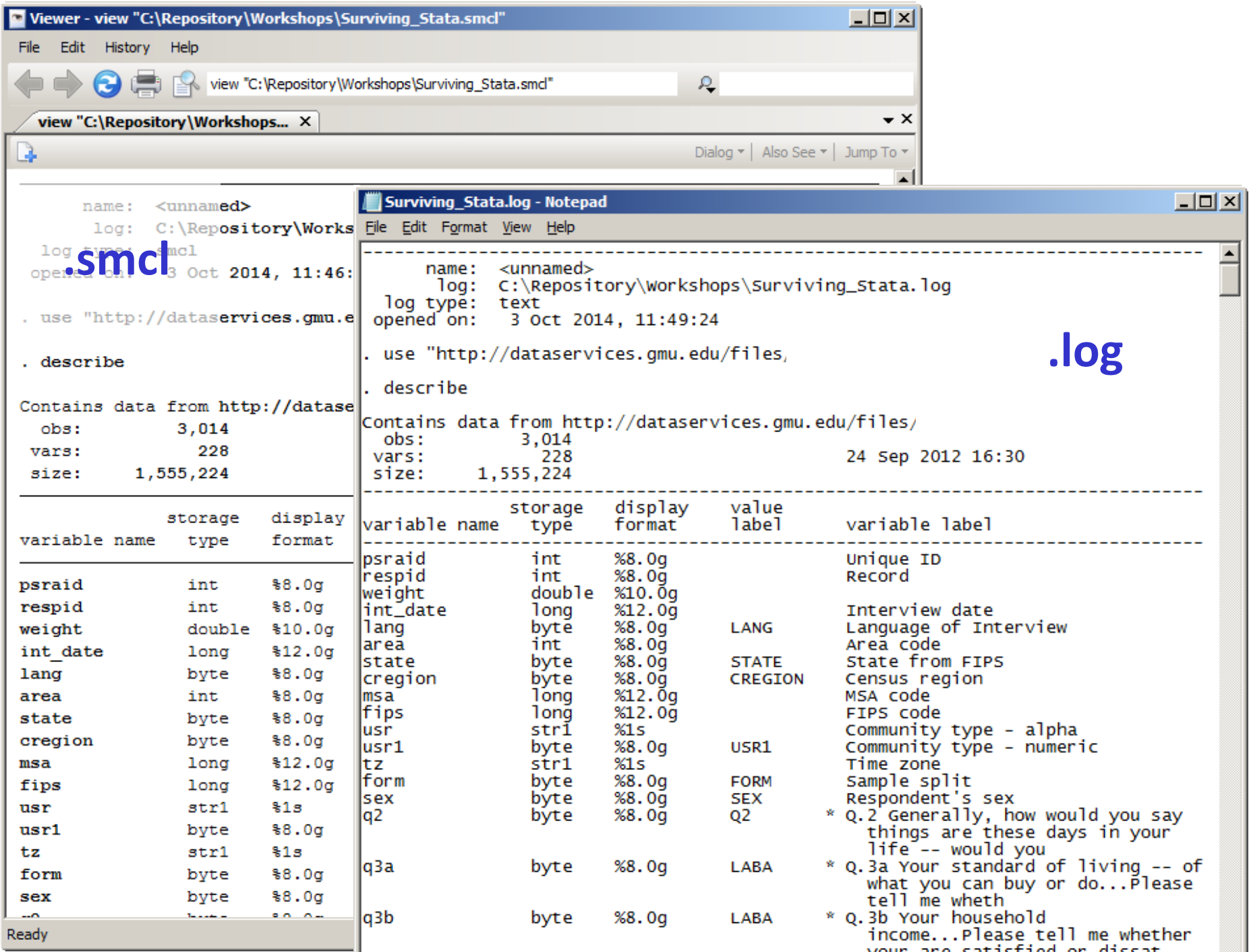
Browse & Edit

Variables Manager
edit labels & properties

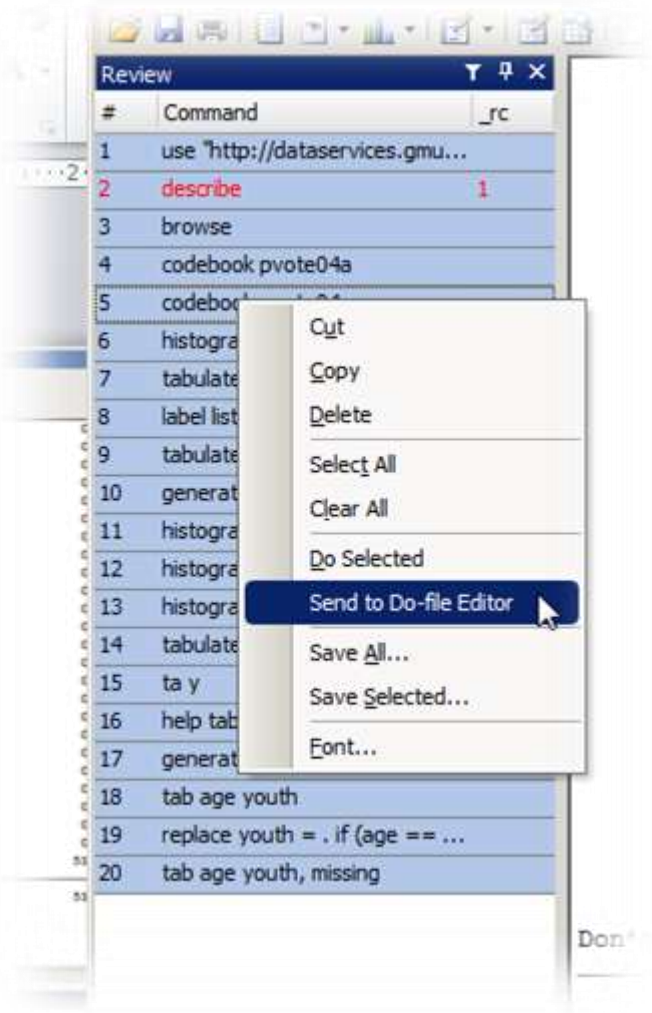
More/ Break

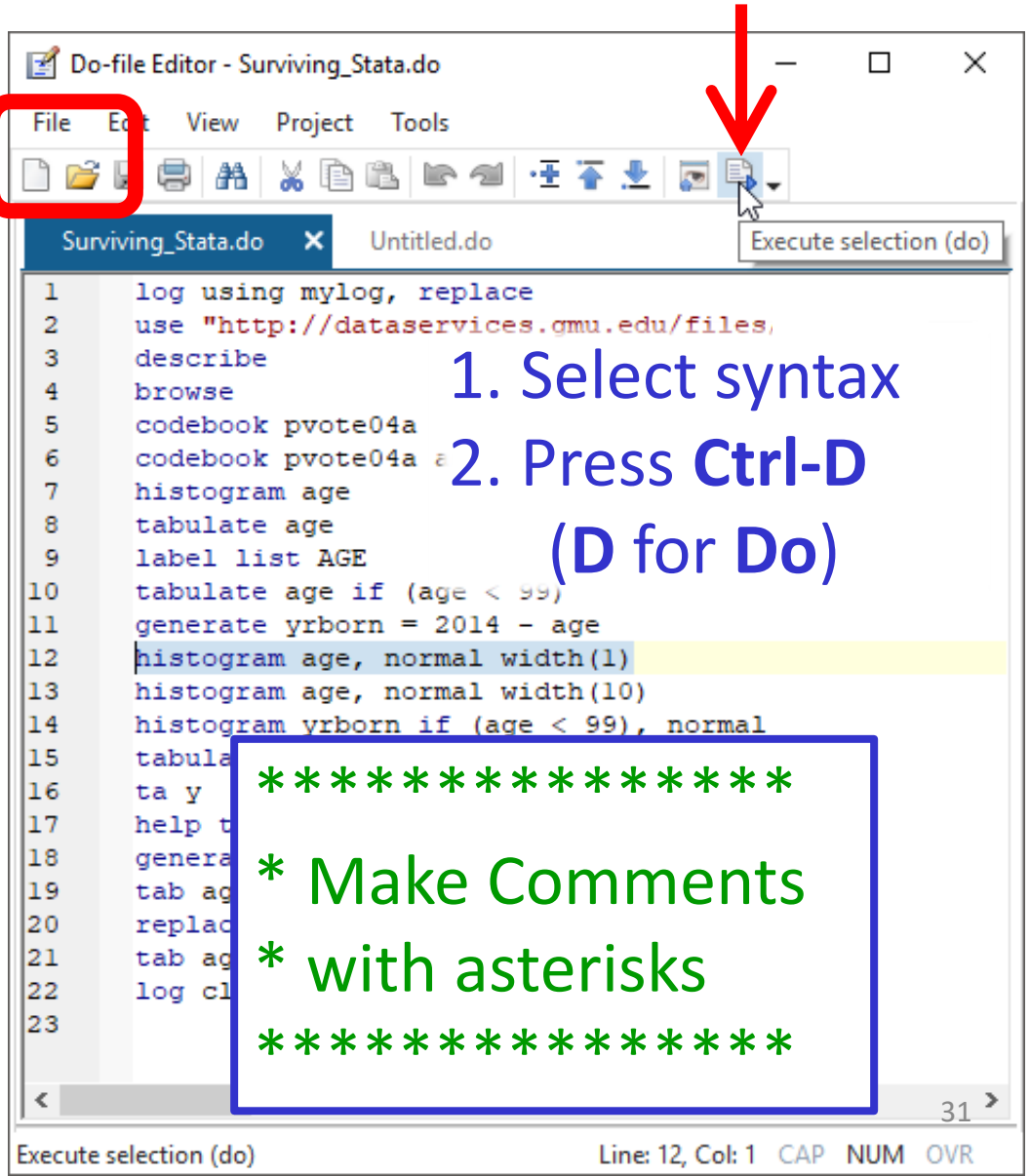
Log File Types

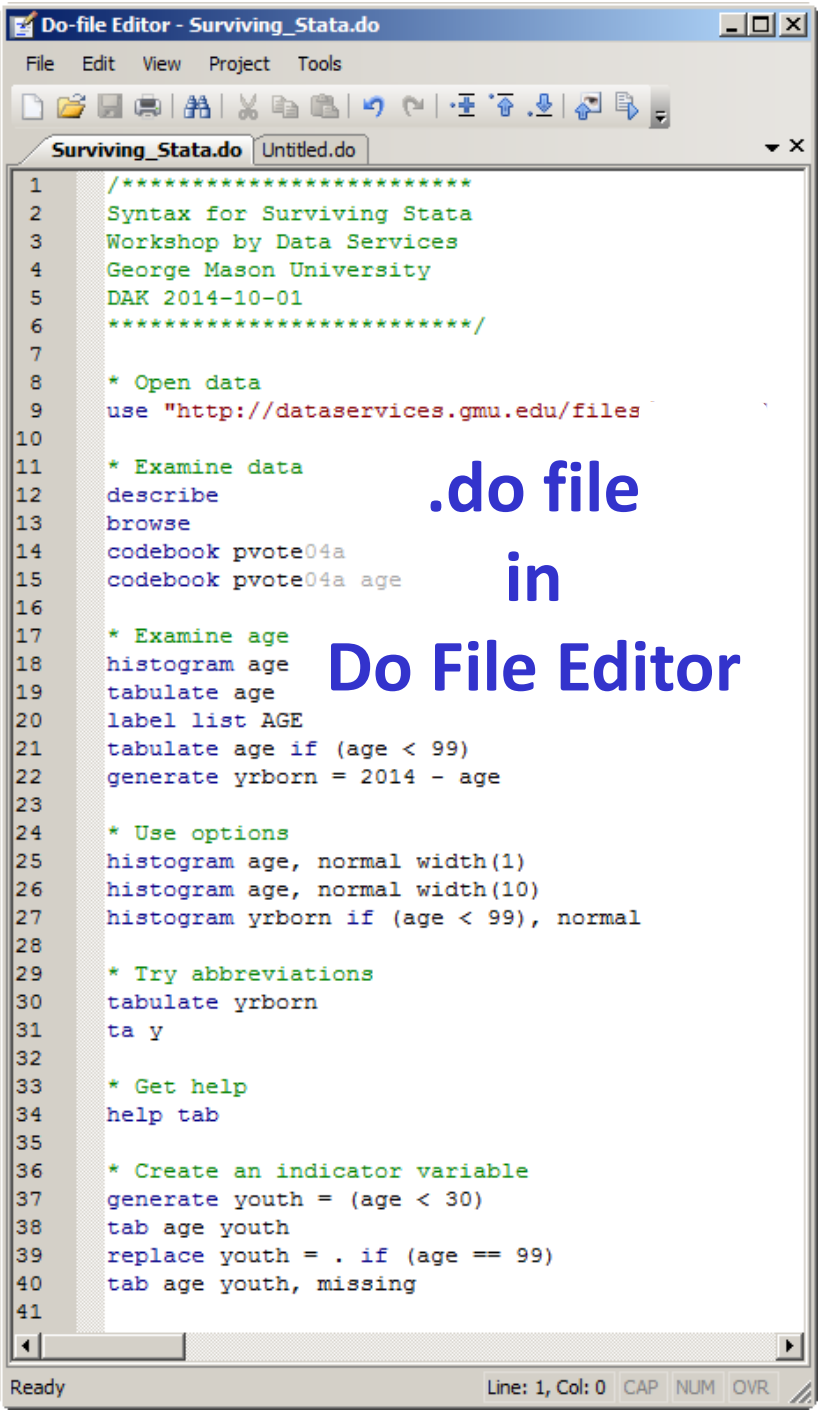



Viewer - view "C:\Repository\Workshops\Surviving_Stata.smcl"
File Edit History Help
view "C:\Repository\Workshops\Surviving_Stata.smcl"
view "C:\Repository\Workshops...
Dialog
Also See
Jump To
name: <unnamed>
log: C:\Repository\Works
log type: smcl
opened on: 3 Oct 2014, 11:46:
. use "http://dataservices.gmu.e
. describe
Contains data from http://datase
obs: 3,014
vars: 228
size: 1,555,224
storage display
variable name type format
psraid int %8.0g
respid int %8.0g
weight double %10.0g
int_date long %12.0g
lang byte %8.0g
area int %8.0g
state byte %8.0g
cregion byte %8.0g
msa long %12.0g
fips long %12.0g
usr str1 %1s
usr1 byte %8.0g
tz str1 %1s
form byte %8.0g
sex byte %8.0g
Ready
.smcl
Surviving_Stata.log - Notepad
File Edit Format View Help
name: <unnamed>
log: C:\Repository\Workshops\Surviving_Stata.log
log type: text
opened on: 3 Oct 2014, 11:49:24
. use "http://dataservices.gmu.edu/files/
. describe
Contains data from http://dataservices.gmu.edu/files/
obs: 3,014
vars: 228 24 Sep 2012 16:30
size: 1,555,224
storage display value
variable name type format label variable label
psraid int %8.0g Unique ID
respid int %8.0g Record
weight double %10.0g
int_date long %12.0g Interview date
lang byte %8.0g LANG Language of Interview
area int %8.0g Area code
state byte %8.0g STATE State from FIPS
cregion byte %8.0g CREGION Census region
msa long %12.0g MSA code
fips long %12.0g FIPS code
usr str1 %1s Community type - alpha
usr1 byte %8.0g USR1 Community type - numeric
tz str1 %1s Time zone
form byte %8.0g FORM Sample split
sex byte %8.0g SEX Respondent's sex
q2 byte %8.0g Q2 * Q.2 Generally, how would you say things are these days in your life -- would you
q3a byte %8.0g LABA * Q.3a Your standard of living -- of what you can buy or do...Please tell me wheth
q3b byte %8.0g LABA * Q.3b Your household income...Please tell me whether
.log

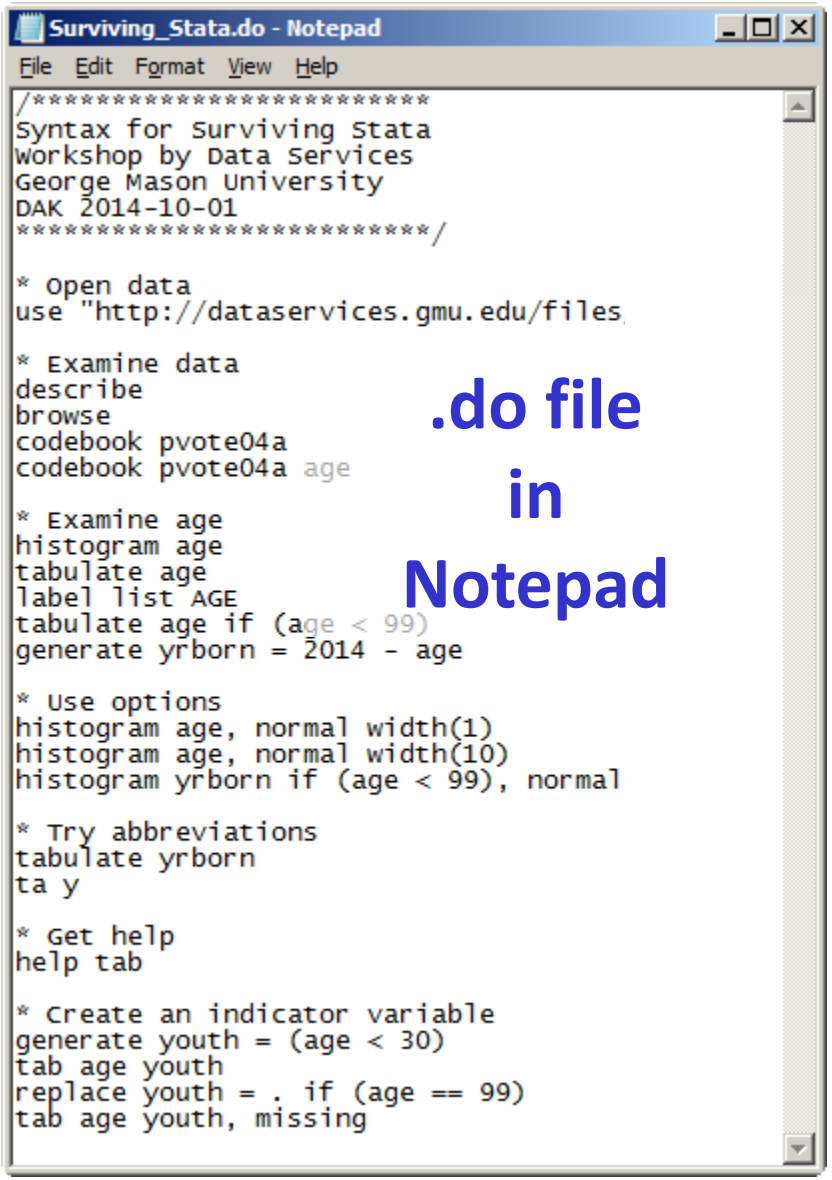
Do- Files


Review
Command _rc
1 use "http://dataservices.gmu...
2 describe 1
3 browse
4 codebook pvote04a
5 codebo
6 histogra
7 tabulate
8 label list
9 tabulate
10 generat
11 histogra
12 histogra
13 histogra
14 tabulate
15 ta y
16 help tab
17 generat
18 tab age youth
19 replace youth = . if (age == ...
20 tab age youth, missing
Cut
Copy
Delete
Select All
Clear All
Do Selected
Send to Do-file Editor
Save All...
Save Selected...
Font...


Do-file Editor - Surviving_Stata.do
File Edit View Project Tools
Execute selection (do)
Surviving_Stata.do
Untitled.do
1 log using mylog, replace
2 use "http://dataservices.gmu.edu/files,
3 describe
4 browse
5 codebook pvote04a
6 codebook pvote04a
7 histogram age
8 tabulate age
9 label list AGE
10 tabulate age if (age < 99)
11 generate yrborn = 2014 - age
12 histogram age, normal width(1)
13 histogram age, normal width(10)
14 histogram yrborn if (age < 99), normal
15 tabula
16 ta y
17 help t
18 genera
19 tab ag
20 replac
21 tab ag
22 log cl
23
1. Select syntax
2. Press Ctrl-D
(D for Do)
****
* Make Comments
* with asterisks
****
31
Execute selection (do)
Line: 12, Col: 1 CAP NUM OVR



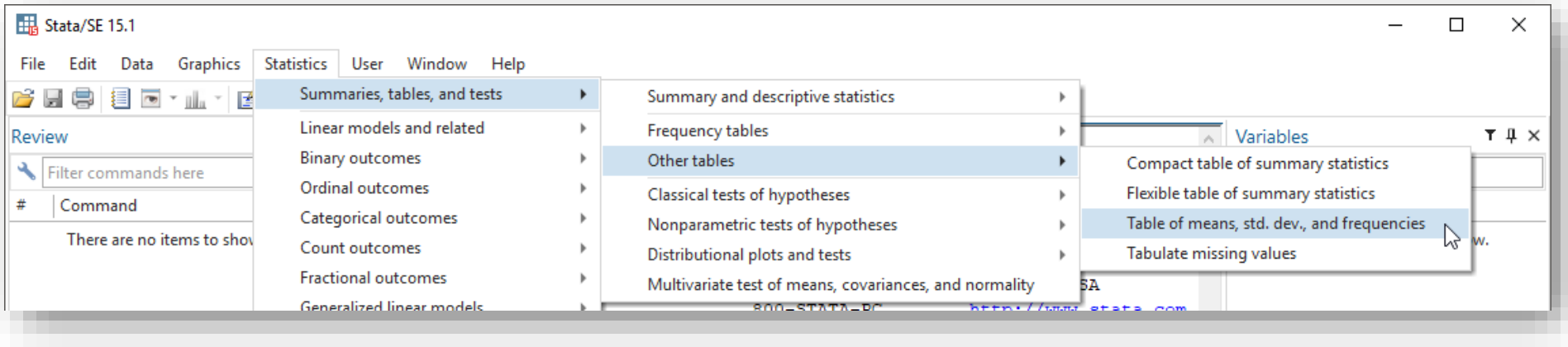
.do file in Do File Editor



.do file in Notepad

Using **help**

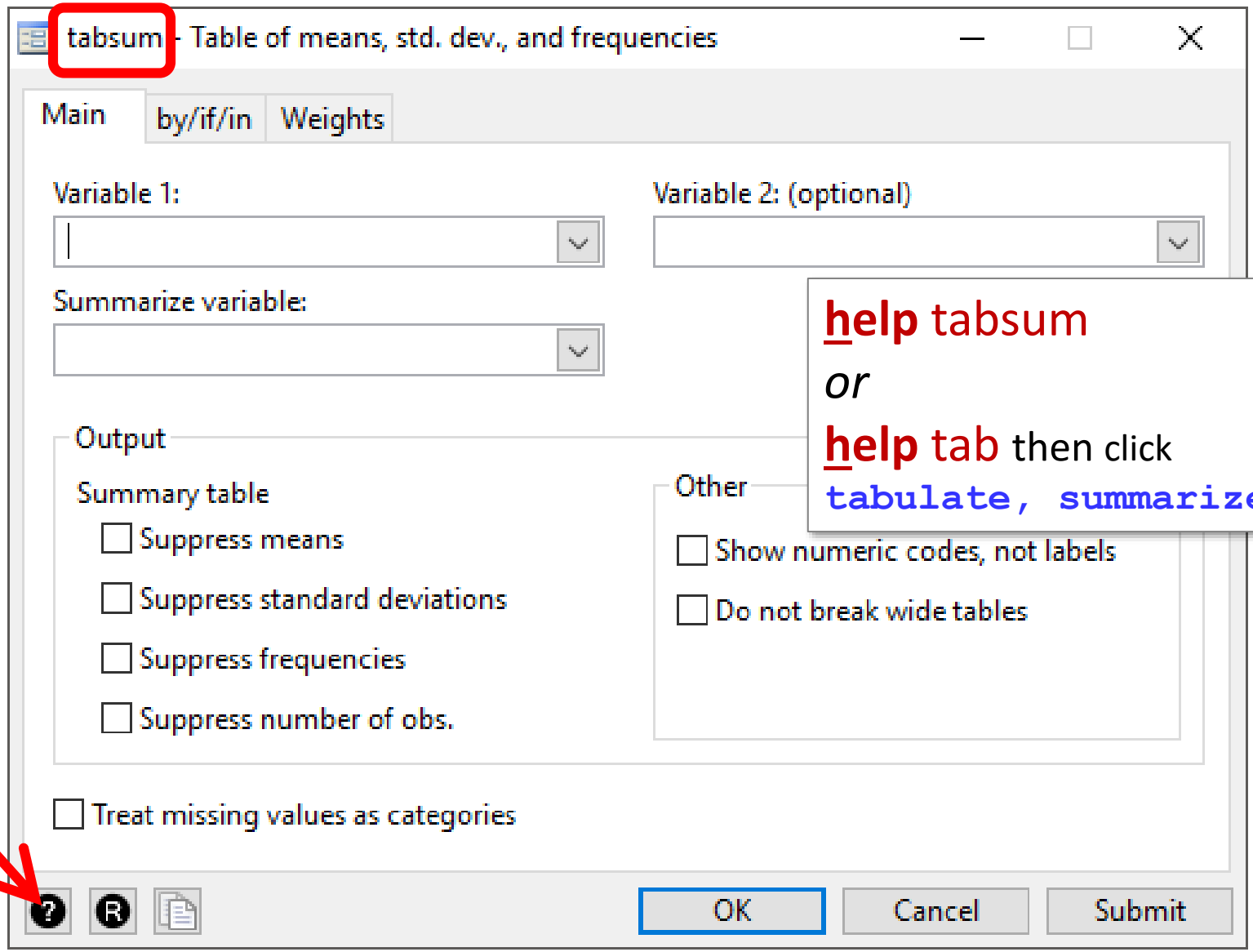
Why Use Syntax?



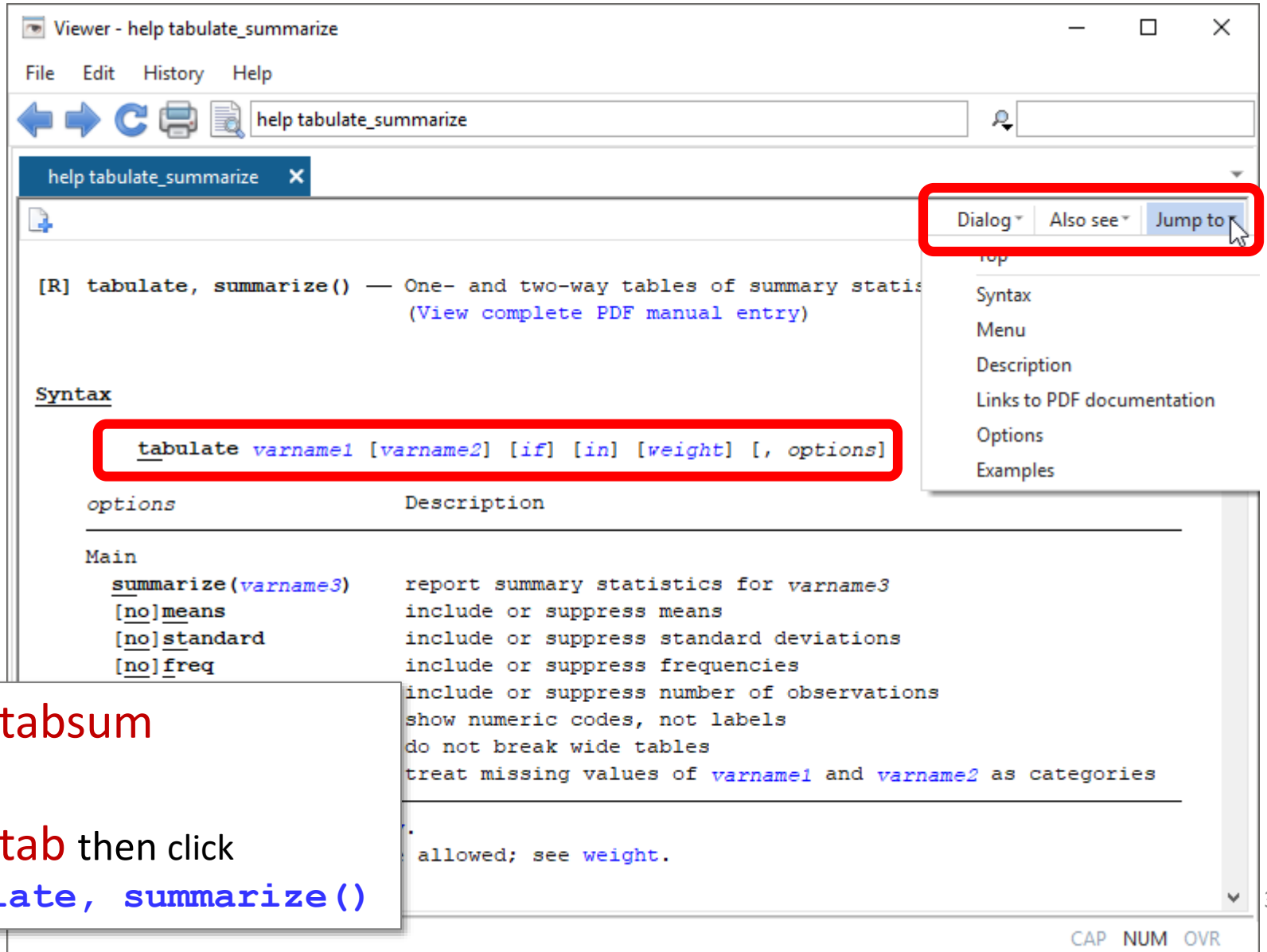
Most people do the same few things over and over again.

log	cd	use	save	clear	drop	keep
recode	encode	generate	egen	replace	rename	
tabulate	summarize	ttest	logistic	regress	estimates	

Getting to Help



The Help Viewer



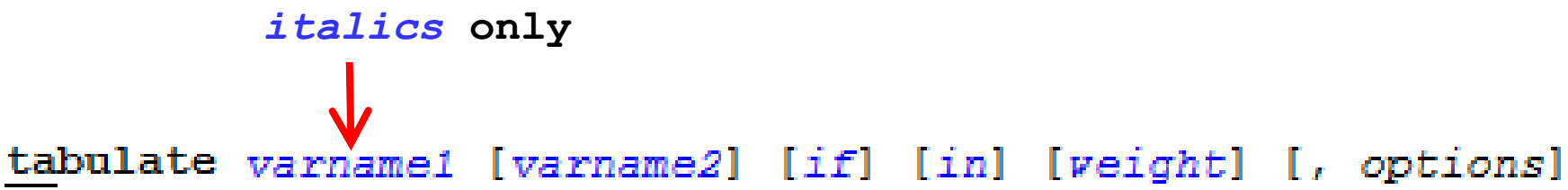
help tabsum

or

help tab then click

tabulate, summarize()

Syntax



Underline **Blue** **[Brackets]** ***Italics***

must type **click here**

replace me **[not needed]**

Simplifying Syntax

Title

[D] generate — Create or change contents of variable

Syntax

Create new variable

generate [*type*] *newvar*[:*lblname*] =*exp* [*if*] [*in*]

generate *newvar =exp*

Practice with Options

help histogram

More Options

[R] histogram — Histograms for continuous and categorical variables
(View complete PDF manual entry)

Syntax

histogram *varname* [*if*] [*in*] [*weight*] [, [*continuous_opts* | *discrete_opts*] *options*]

continuous_opts

Main
- bin(#)
- width(#)
- start(#)

discrete_opts

Main
- discrete
- width(#)
- start(#)

options

options	Description
Main	
density	draw as density; the defaul
fraction	draw as fractions
frequency	draw as frequencies
percent	draw as percentages
bar_options	rendition of bars
binrescale	recalculate bin sizes when
addlabels	add height labels to bars
addlabopts(*marker_label_options*)	affect rendition of labels
Density plots	
normal	add a normal density to the
normopts(*line_options*)	affect rendition of normal
kdensity	add a kernel density estima
kdenopts(*kdensity options*)	affect rendition of kernel

Quick Start - NEW!

histogram — Histograms for continuous and categorical variables

Description

`histogram` draws histograms of *varname*, which is assumed to be the name of a continuous variable unless the `discrete` option is specified.

`hist` is a synonym for `histogram`.

Quick start

Histogram of `v1`

```
histogram v1
```

Add a normal density curve to the graph

```
histogram v1, normal
```

Add a kernal density estimate to the graph

```
histogram v1, normal kdensity
```

Add "My Title" as the title of the graph

```
histogram v1, normal kdensity title("My Title")
```

Specify the number of bins as 10

```
histogram v1, bin(10)
```

Specify the width of the bins as 2

```
histogram v1, width(2)
```

Specify that `v2` should be treated as discrete

```
histogram v2, discrete
```

More on Syntax

Great Tutorial

https://www.ssc.wisc.edu/sscc/pubs/stat.htm

Stata Basic Syntax

command varlist qualifiers, options

command	1, 2, or 3 words that indicate what to do
varlist	0, 1, or more variables to work with
qualifiers	add labels or limits to the instructions
, *options*	specify additions or changes to the output

Legend
Replace any syntax in *italics* with your own variable names or values.

var	any variable ("var")
cat	a categorical (nominal) var
num	a numeric (interval/scale) var
x	the predictor or Independent var
y	the response or Dependent var

Describe Values

ANY VARIABLE(S)

```
codebook vars
```

CATEGORICAL

```
tabulate cat
graph bar, over(cat)
  also graph pie, ...
```

NUMERIC

```
summarize num
mean num
histogram num, freq
  if integer, add discrete
```

MORE OPTIONS

Install once before use.

```
ssc install fre
fre cat
ssc install univar
univar num
```

Show Relationships

TWO NUMERIC

```
pwcorr num num
scatter y x || lfit y x
```

TWO CATEGORICAL

```
tabulate x y, row
tabulate y x, col
```

NUMERIC + CATEGORICAL

```
tabulate cat, sum(num)
table cat, c(mean num)
graph bar num, over(cat)
    also graph box ...
```

3-WAY RELATIONSHIPS

```
table cat cat, by(cat)
table cat cat, c(mean num)
scatter num num, by(cat)
scatter num num [w=num]
```

Data Management

SIMPLIFY THE FILE

```
keep var1 var2 etc
keep if var != 4
```

CHANGE VALUES

```
generate newvar = var
  ↑ avoid changing original var
recode var (2 3 = 1)
recode var (99 = .)
        missing value ↑
replace var = 1 if var == 9
replace var = "US"
        if var == "VA"
    strings must be quoted ↑
```

Modifiers

Add to other commands

IF EXPRESSION

Include obs only if true

```
tab var if var == 4
```

BY PREFIX COMMAND

Do for (by) each value in *cat*

```
bysort cat : tab var
```

OPERATORS

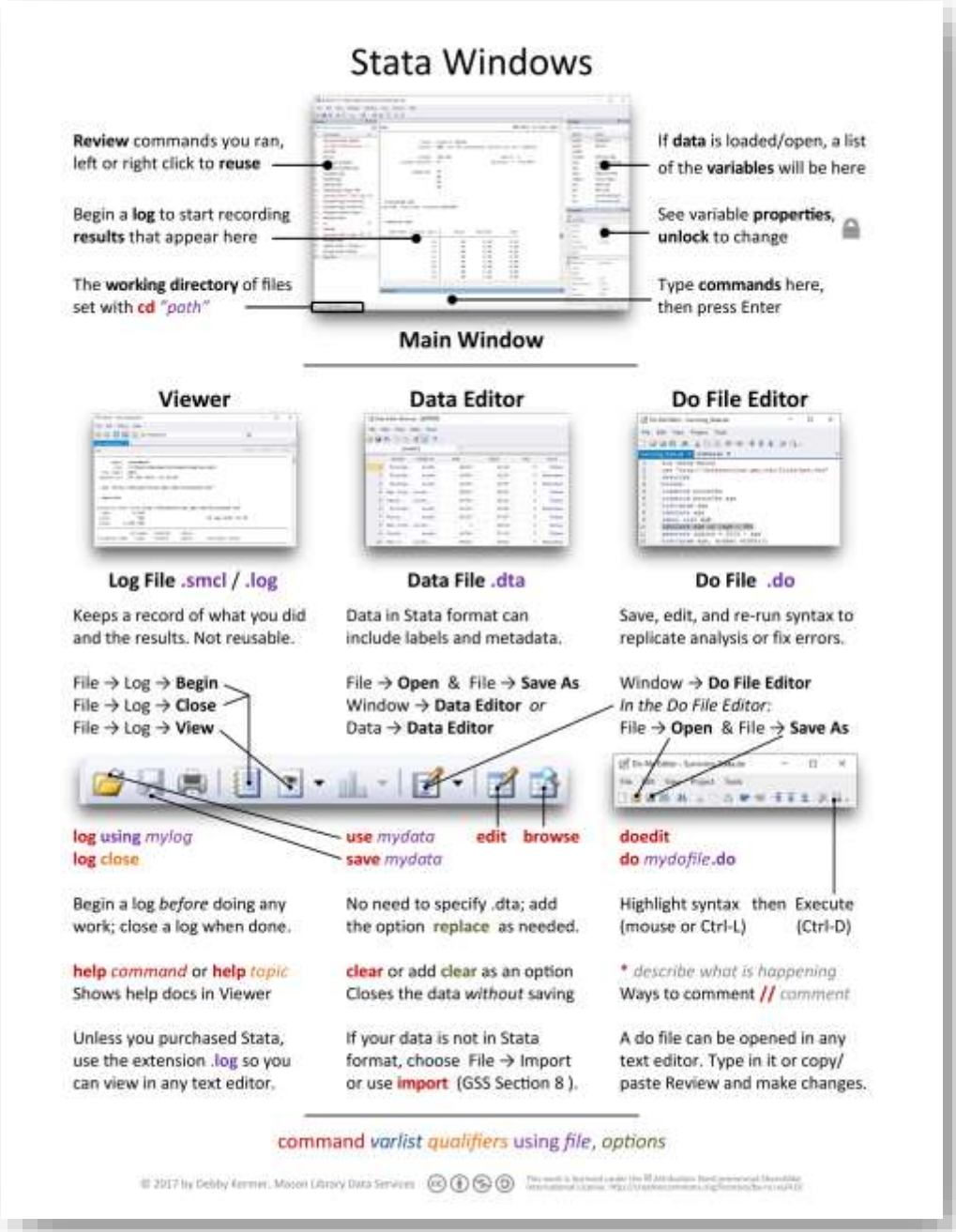
== Equal to	< Less Than
!= Not Equal to	<= Less than or Equal to
& AND	| OR

Significance Tests

TEST	SYNTAX	P-VALUE
Chi-Square	`tabulate cat cat, chi2`	after `Pr =`
Correlation	`pwcorr num num, sig`	under Pearson's r
One-Way ANOVA	`oneway num cat, tabulate`	under `Prob > F`

see the bottom of **http://infoguides.gmu.edu/software/stata**

Save & Close



see the bottom of **http://infoguides.gmu.edu/software/stata**

Review

Review

What are the **4 parts** of Stata Syntax, in order?

command varlist qualifiers, options

What do **[**brackets**]** mean in Stata Syntax Help?

not needed / optional element

What does *italics* mean in Stata Syntax Help?

replace with your own word(s) or value(s)

Syntax

command varlist qualifiers, options

command	1, 2, or 3 words specifying the task
varlist	0, 1, 2 or more variables
= *exp*	A math or logical statement to set a value
if *exp*	A logical statement to limit cases
, *options*	A word or selector to alter the command

Qualifiers

=exp expression (math or logical)

```
generate yrborn    = 2005 - age
generate youth     = ( age < 30 )
```

[*if*] logical statement

```
histogram age if ( age != 99 )
replace youth = . if ( age == 99 )
```

[*in*] range of case numbers

```
tabulate pvote04a in 1/100
```

… names or labels

```
label yrborn "Year of Birth"
label list AGE
```

Logical Operators	
|	or
&	and
==	is equal to
!=	not equal to
<	less than
<=	less than or equal to

End

For these workshop **slides**:

https://**dsc**.gmu.edu/**workshops/stata**

For more resources on **Stata**:

https://**infoguides**.gmu.edu/**software/stata**

For help with **511**:

https://**infoguides**.gmu.edu/**statsclass**

To provide **feedback** about this workshop:

https://**dsc**.gmu.edu/**workshops/feedback**

Debby Kermer dkermer@gmu.edu