

Introduction to Stata

Leonardo Collado-Torres

December 26, 2016

- Have you heard about Stata?

What is Stata?

- Statistical software
- Has a user interface *and* code
- Most statistical methods are implemented
- It's licensed software, so it's not free

Why Stata?

- User interface: means users do not need to know how to write code
- It is great for mixed models, survey data analysis, etc
- Can be used for reproducible work via *do* files
- Researchers can contribute their modules: there's even a Stata journal

Main components

- Data browser
 - View your data
 - *Can* edit but not recommended
- Help files: very detailed and interconnected
- Can import data from many file types
- Intuitive menu
- Console: shows the latest code
- Log: shows the latest results
- *do* file editor: make your work reproducible

Are you ready to start using Stata?

First, type the following command (or copy paste it)

```
sysuse autos
```

What do you get?

```
sysuse autos
```

```
. sysuse autos
file "autos.dta" not found
r(601);

end of do-file
r(601);
```

So, what went wrong?

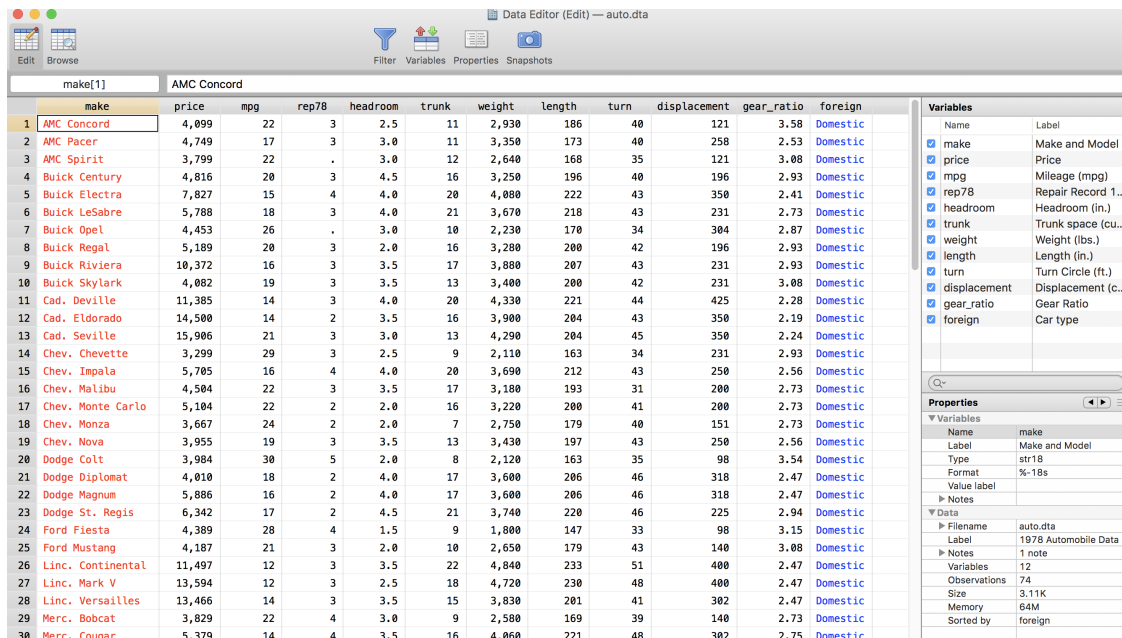
Lets see if the error message gives us a hint as to what could have gone wrong.

The filename you specified cannot be found. Perhaps you mistyped the name, or it may be on another CDor directory.

Ohh, it could be that we incorrectly typed the name! Try this next:

```
sysuse auto
```

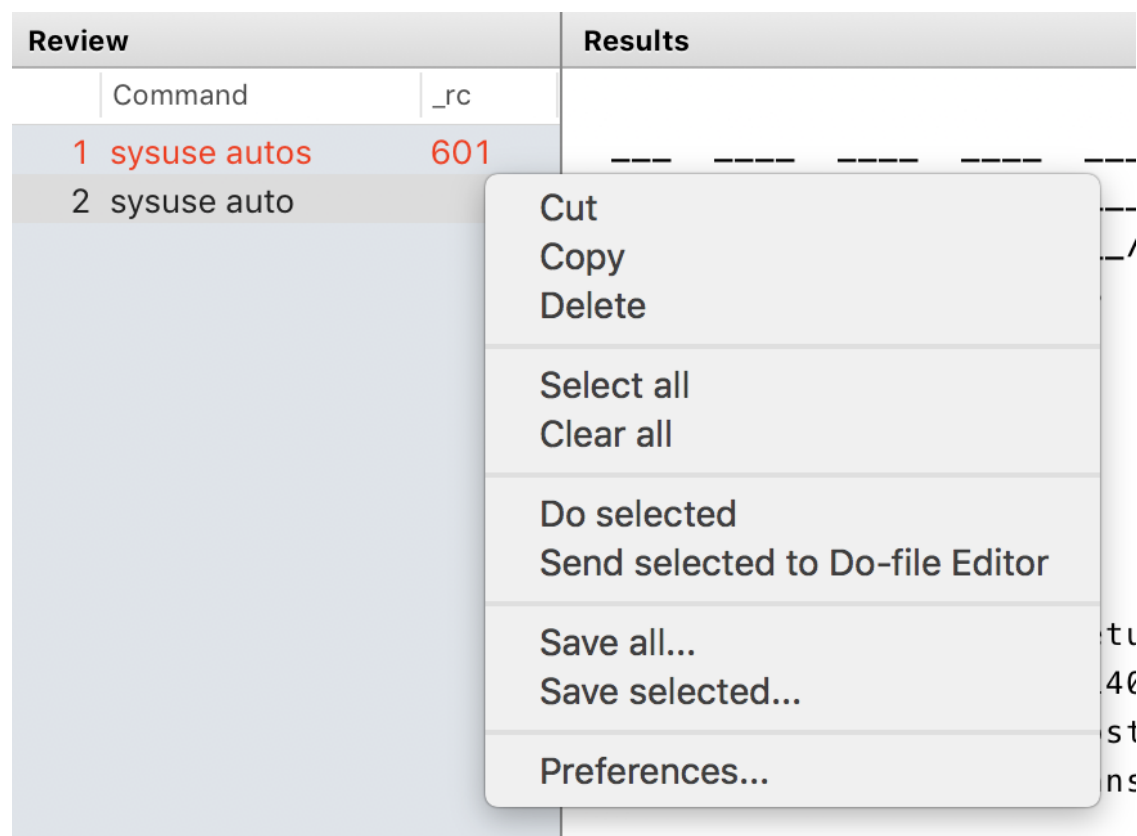
Now lets explore the data



	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.47	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15	Domestic
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179	43	140	3.08	Domestic
26	Linc. Continental	11,497	12	3	3.5	22	4,840	233	51	400	2.47	Domestic
27	Linc. Mark V	13,594	12	3	2.5	18	4,720	230	48	400	2.47	Domestic
28	Linc. Versailles	13,466	14	3	3.5	15	3,830	201	41	302	2.47	Domestic
29	Merc. Bobcat	3,829	22	4	3.0	9	2,580	169	39	140	2.73	Domestic
30	Merc. Cougar	5,370	14	4	3.5	16	4,860	221	48	302	2.75	Domestic

Command review

Review			Results
	Command	_rc	
1	sysuse autos	601	
2	sysuse auto		

A screenshot of the Stata Command Review window. The window has a table with columns 'Review', 'Command', '_rc', and 'Results'. The first row shows command '1 sysuse autos' with result '601'. The second row shows '2 sysuse auto'. A context menu is open over the table, showing options: Cut, Copy, Delete, Select all, Clear all, Do selected, Send selected to Do-file Editor, Save all..., Save selected..., and Preferences....

Basic syntax

What do you want to do?

- Described by an action
- Think of it as a **verb**
- For example, load a data set included in Stata with **sysuse**

What are you going to use for that action?

- That's the actual data that you will use
- Think of it as the **subject**
- For example, the **auto** data set

Keep track of your work

- Do you remember all the options in a user interface you used 6 months ago?
- It's important to keep track of your work!
- Useful when learning so you can revisit what you did: see what worked, what didn't
- Two main options in Stata: log files and do files

Log files

These files keep track of everything:

- The commands you used
- The commands that specify what you did with the user interface
- The output you generated: tables, results
- They do not save images

Start a log for our session

Example log file

```

name: <unnamed>
log: /Users/lcollado/Desktop/prueba.smcl
log type: smcl
opened on: 23 Dec 2016, 17:39:21

. tab foreign

      Car type |      Freq.   Percent   Cum.
-----+-----
      Domestic |          52    70.27    70.27
      Foreign  |          22    29.73   100.00
-----+-----
          Total |          74   100.00

. graph bar foreign

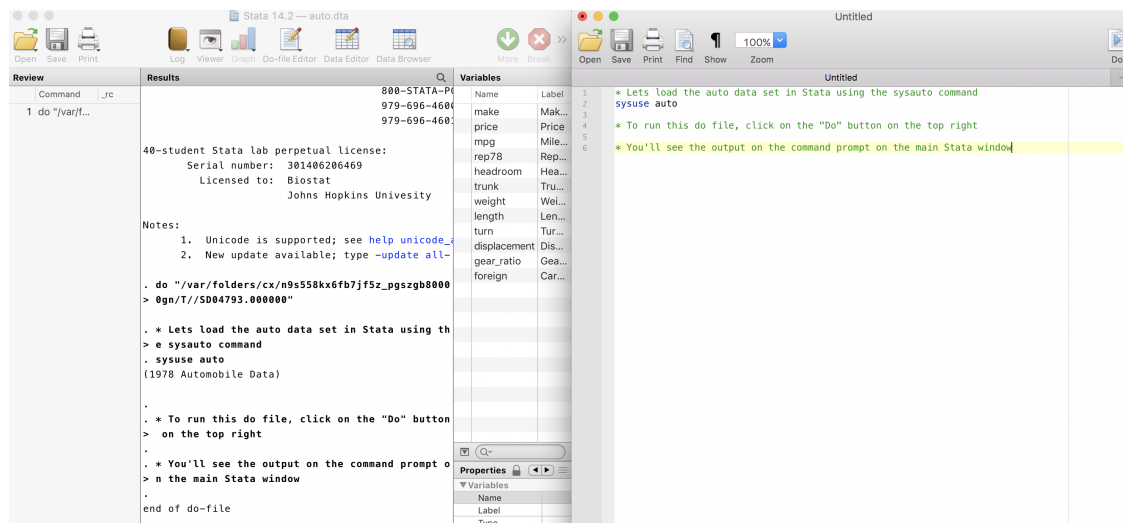
. exit

```

Do files

- These files contain commands only
- They are an executive summary of what you did
- Great for analyses you want to share
- You can include comments describing the logic of what you are doing
- You can *execute* them to run commands
- Cleaner than *log* files.
- Open the do file and paste the command to load the `auto` data set.

Example do file



Exercise

Use the Afghanistan data from the Worldbank and make a plot of the agricultural land (in squared kilometers). Start by loading the data!

No cheating!!!

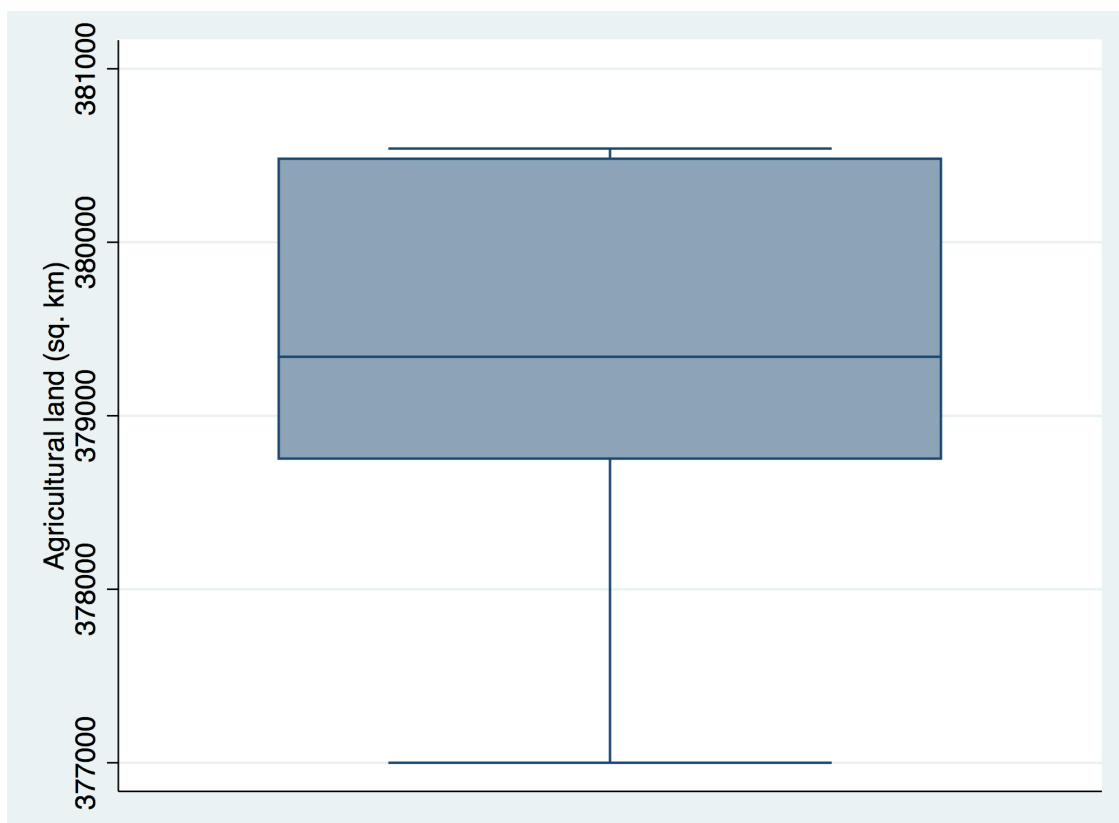
Don't look at the next slides! Try first!

What variable is it?

Variables		
	Name	Label
	AGAGRTRACNO	Agricultural machinery, tractors
	AGLNDAGRIK2	Agricultural land (sq. km)
	AGLNDAGRIZS	Agricultural land (% of land area)
	AGLNDIRIGAGZS	Agricultural irrigated land (% of total agricultural land)
	AGLNDTRACZS	Agricultural machinery, tractors per 100 sq. km of arable land
	ENATMMETHAGKTCE	Agricultural methane emissions (thousand metric tons of CO2 equivalent)
	ENATMMETHAGZS	Agricultural methane emissions (% of total)
	ENATMNOXEAGKTCE	Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)
	ENATMNOXEAGZS	Agricultural nitrous oxide emissions (% of total)
	SLEMPINSVFEZS	Share of women in wage employment in the nonagricultural sector (% of total nona
	SLISVIFRMFEZS	Informal employment, female (% of total non-agricultural employment)
	SLISVIFRMMAZS	Informal employment, male (% of total non-agricultural employment)
	SLISVIFRMZS	Informal employment (% of total non-agricultural employment)
	TMVALAGRIZSUN	Agricultural raw materials imports (% of merchandise imports)
	TXVALAGRIZSUN	Agricultural raw materials exports (% of merchandise exports)

It's AGLNDAGRIK2

```
use arg_worldbank_2016.dta, clear
graph box AGLNDAGRIK2
```



Looks pretty easy, right? Just 2 lines of code.

Data Source									
A	B	C	D	E	F	G	H	I	
1	Data Source	World Development Indicators							
2	Last Updated Date	12/16/16							
3									
4	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964
5	Afghanistan	AFG	Mobile account, income, richest 60% (% age 15+) [w2]	WP15163_4.9					
6	Afghanistan	AFG	Mobile account, income, poorest 40% (% age 15+) [w2]	WP15163_4.8					
7	Afghanistan	AFG	Mobile account, female (% age 15+) [w2]	WP15163_4.3					
8	Afghanistan	AFG	Mobile account, male (% age 15+) [w2]	WP15163_4.2					
9	Afghanistan	AFG	Mobile account (% age 15+) [w2]	WP15163_4.1					
10	Afghanistan	AFG	Account at a financial institution, income, richest 60% (% age 15+) [ts]	WP_time_01.9					
11	Afghanistan	AFG	Account at a financial institution, income, poorest 40% (% age 15+) [ts]	WP_time_01.8					
12	Afghanistan	AFG	Account at a financial institution, female (% age 15+) [ts]	WP_time_01.3					
13	Afghanistan	AFG	Account at a financial institution, male (% age 15+) [ts]	WP_time_01.2					
14	Afghanistan	AFG	Account at a financial institution (% age 15+) [ts]	WP_time_01.1					
15	Afghanistan	AFG	Presence of peace keepers (number of troops, police, and military c	VC.PKP.TOTL.UN					
16	Afghanistan	AFG	Intentional homicides (per 100,000 people)	VC.IHR.PSRC.P5					
17	Afghanistan	AFG	Internally displaced persons (number, low estimate)	VC.IDP.TOTL.LE					
18	Afghanistan	AFG	Internally displaced persons (number, high estimate)	VC.IDP.TOTL.HE					
19	Afghanistan	AFG	Battle-related deaths (number of people)	VC.BTL.DETH					
20	Afghanistan	AFG	Travel services (% of commercial service exports)	TX.VAL.TRVL.ZS.WT					
21	Afghanistan	AFG	Transport services (% of commercial service exports)	TX.VAL.TRVL.ZS.WT					
22	Afghanistan	AFG	High-technology exports (% of manufactured exports)	TX.VAL.TECH.MF.ZS					
23	Afghanistan	AFG	High-technology exports (current US\$)	TX.VAL.TECH.CD					
24	Afghanistan	AFG	Commercial service exports (current US\$)	TX.VAL.SERV.CD.WT					
25	Afghanistan	AFG	Computer, communications and other services (% of commercial se	TX.VAL.OTHER.ZS.WT					
26	Afghanistan	AFG	Export value index (2000 = 100)	TX.VAL.MRCH.XD.WD					
27	Afghanistan	AFG	Merchandise exports to low- and middle-income economies within	TX.VAL.MRCH.WR.ZS	23.24649299	12.35955056	13.58234295	18.54493581	25.43103448
28	Afghanistan	AFG	Merchandise exports by the reporting economy (current US\$)	TX.VAL.MRCH.WL.CD	49900000	53400000	58900000	70100000	69600000
29	Afghanistan	AFG	Merchandise exports by the reporting economy, residual (% of total	TX.VAL.MRCH.RS.ZS	30.26052104	42.13483146	51.10356537	46.07703281	44.25287356
30	Afghanistan	AFG	Merchandise exports to low- and middle-income economies in Sub-	TX.VAL.MRCH.R6.ZS					

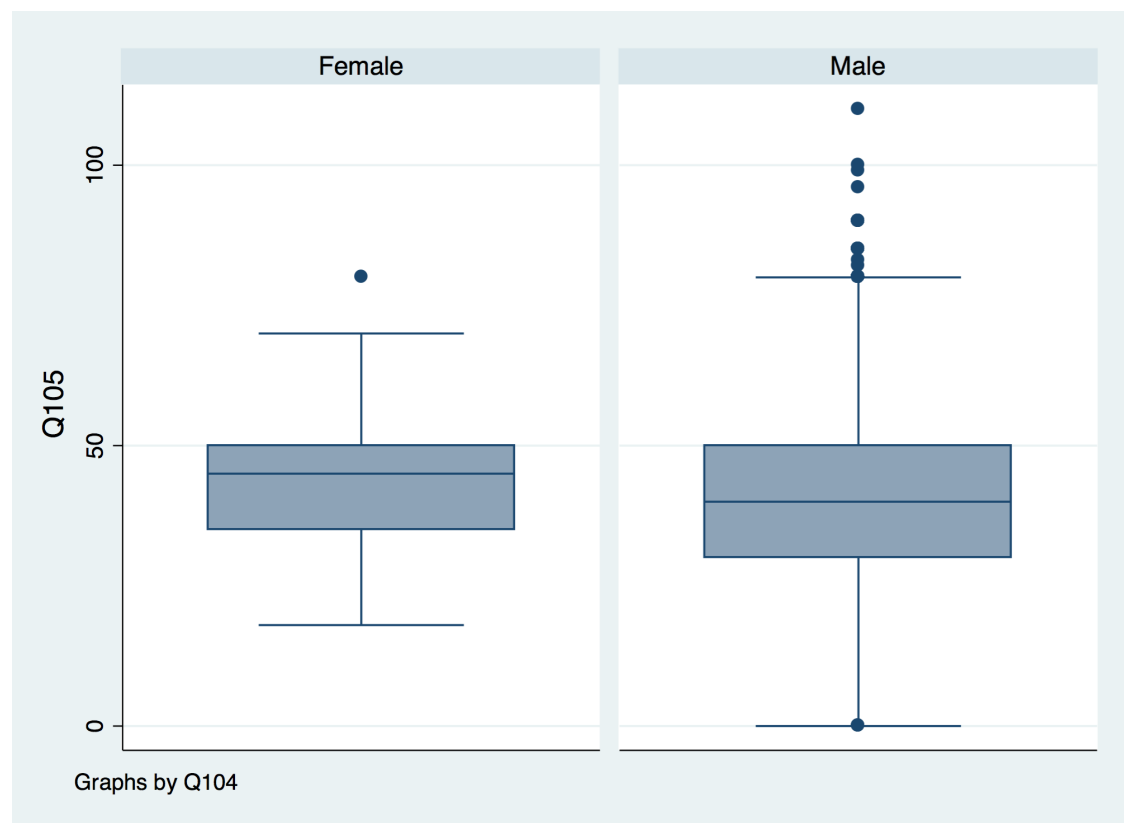
The original data was a bit more messy! Transforming it to something we can use is called *data cleaning* or *data tidying*.

Stata syntax: options

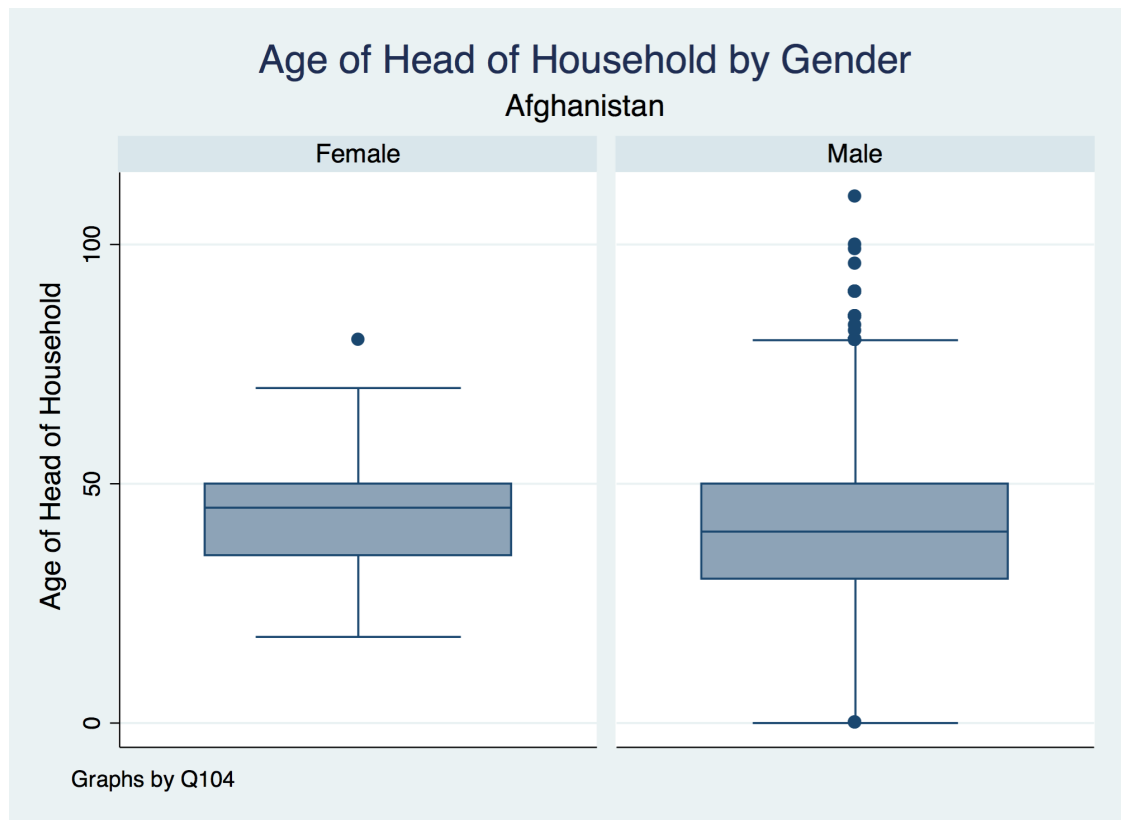
- Many Stata commands have options that give us more fine tuned control over what we want to do
- You can find them also via the user interface: many are checkboxes
- Think of options as **adjectives**
- Examples: title of a graph, which data to use, colors
- They normally come after a comma
- You can find the options in the Stata help

Options example

```
use "HH Listing.dta"  
graph box q105 if q103 == 1, by(q104)
```



```
graph box q105 if q103 == 1, ytitle(Age of Head of Household) /*  
    */ by(, title(Age of Head of Household by Gender) /*  
    */ subtitle(Afghanistan)) by(q104)
```



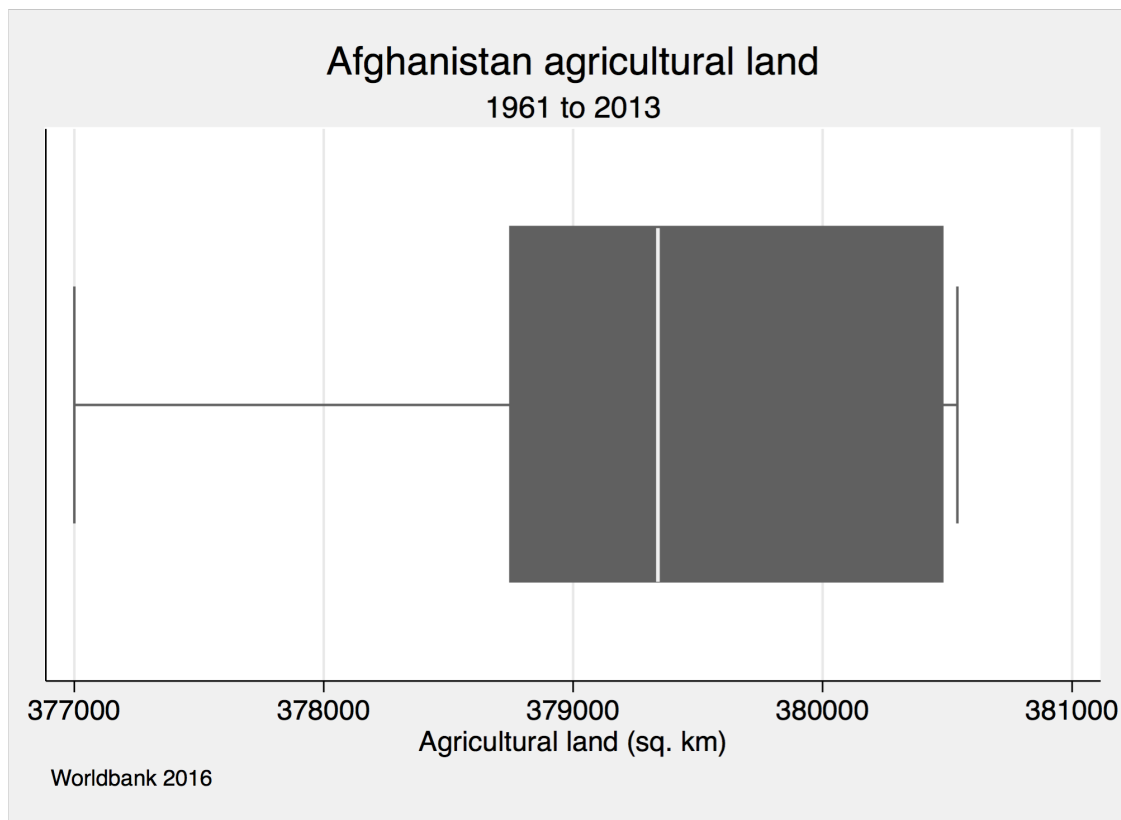
Stata help

- It comes in two shapes: help pages for commands and PDF documents
- Help pages are more direct
- Sometimes it's easier to scroll down to the examples section
- The help pages are interconnected, so use the hyperlinks
- Check out the help page for graph box: `help graph box`

Exercise continued

Improve the plot you made earlier as if you were to show the distribution of the agricultural land of Afghanistan in the media (news) or a journal.

```
graph hbox AGLNDAGRIK2, title(Afghanistan agricultural land) /*
    */ subtitle(1961 to 2013) note(Worldbank 2016) scheme(sj)
```

See all the entries for a variable: `list`

```
list AGLNDAGRIK2
```

```
. use arg_worldbank_2016.dta, clear
(Written by R. )
```

```
. list AGLNDAGRIK2
```

```

+-----+
| AGLNDA~2 |
+-----+
1. |      . |
2. |  377000 |
3. |  377600 |
4. |  378100 |
5. |  378730 |
+-----+
6. |  378750 |
7. |  379130 |
8. |  379790 |
9. |  379800 |
10. |  379960 |
+-----+
11. |  380060 |
```

12.		380360	
13.		380460	
14.		380480	
15.		380480	

16.		380480	
17.		380480	
18.		380500	
19.		380500	
20.		380490	

21.		380490	
22.		380530	
23.		380540	
24.		380540	
25.		380540	

26.		380540	
27.		380540	
28.		380450	
29.		380400	
30.		380400	

31.		380400	
32.		380300	
33.		380300	
34.		379340	
35.		378130	

36.		377530	
37.		377520	
38.		377900	
39.		378670	
40.		377530	

41.		377530	
42.		377530	
43.		377530	
44.		379100	
45.		379110	

46.		379100	
47.		379100	
48.		379100	
49.		379100	
50.		379100	

51.		379100	
52.		379100	
53.		379100	
54.		379100	
55.		.	

56.		.	

```
57. |          . |
    +-----+
```

codebook: main information

```
codebook AGLNDAGRIK2
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.          )
```

```
. codebook AGLNDAGRIK2
```

```
-----
AGLNDAGRIK2                                     Agricultural
-----
```

```

              type:  numeric (double)
              range:  [377000,380540]          units:  10
unique values:  28                                missing .:  4/57

              mean:    379405
              std. dev: 1110.28

percentiles:      10%      25%      50%      75%      90%
                  377530   378750   379340   380480   380530
```

summarize: univariate statistical summary

```
summarize AGLNDAGRIK2
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.          )
```

```
. summarize AGLNDAGRIK2
```

```

Variable |          Obs          Mean    Std. Dev.        Min        Max
-----+-----
AGLNDAGRIK2 |          53    379404.5    1110.278        377000        380540
```

```
summarize AGLNDAGRIK2, detail
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.          )
```

```
. summarize AGLNDAGRIK2, detail
```

Agricultural land (sq. km)				

	Percentiles	Smallest		
1%	377000	377000		
5%	377530	377520		
10%	377530	377530	Obs	53
25%	378750	377530	Sum of Wgt.	53
50%	379340		Mean	379404.5
		Largest	Std. Dev.	1110.278
75%	380480	380540		
90%	380530	380540	Variance	1232718
95%	380540	380540	Skewness	-.5865218
99%	380540	380540	Kurtosis	2.045356

describe: similar to codebook

```
describe AGLNDAGRIK2
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.          )
```

```
. describe AGLNDAGRIK2
```

	storage	display	value	
variable name	type	format	label	variable label

AGLNDAGRIK2	double	%9.0g		Agricultural land (sq. km)

tabulate: make a table

```
tabulate AGLNDAGRIK2
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.          )
```

```
. tabulate AGLNDAGRIK2
```

Agricultura				
l land (sq.				
km)	Freq.	Percent	Cum.	

377000	1	1.89	1.89	
377520	1	1.89	3.77	
377530	5	9.43	13.21	
377600	1	1.89	15.09	
377900	1	1.89	16.98	

378100		1	1.89	18.87
378130		1	1.89	20.75
378670		1	1.89	22.64
378730		1	1.89	24.53
378750		1	1.89	26.42
379100		10	18.87	45.28
379110		1	1.89	47.17
379130		1	1.89	49.06
379340		1	1.89	50.94
379790		1	1.89	52.83
379800		1	1.89	54.72
379960		1	1.89	56.60
380060		1	1.89	58.49
380300		2	3.77	62.26
380360		1	1.89	64.15
380400		3	5.66	69.81
380450		1	1.89	71.70
380460		1	1.89	73.58
380480		4	7.55	81.13
380490		2	3.77	84.91
380500		2	3.77	88.68
380530		1	1.89	90.57
380540		5	9.43	100.00
-----+-----				
Total		53	100.00	

tabulate: include missing observations

```
tabulate AGLNDAGRIK2, missing
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.)
```

```
. tabulate AGLNDAGRIK2, missing
```

Agricultura				
l land (sq.				
km)		Freq.	Percent	Cum.
-----+-----				
377000		1	1.75	1.75
377520		1	1.75	3.51
377530		5	8.77	12.28
377600		1	1.75	14.04
377900		1	1.75	15.79
378100		1	1.75	17.54
378130		1	1.75	19.30
378670		1	1.75	21.05
378730		1	1.75	22.81
378750		1	1.75	24.56
379100		10	17.54	42.11
379110		1	1.75	43.86
379130		1	1.75	45.61

379340		1	1.75	47.37
379790		1	1.75	49.12
379800		1	1.75	50.88
379960		1	1.75	52.63
380060		1	1.75	54.39
380300		2	3.51	57.89
380360		1	1.75	59.65
380400		3	5.26	64.91
380450		1	1.75	66.67
380460		1	1.75	68.42
380480		4	7.02	75.44
380490		2	3.51	78.95
380500		2	3.51	82.46
380530		1	1.75	84.21
380540		5	8.77	92.98
.		4	7.02	100.00
-----+-----				
Total		57	100.00	

count number of observations

```
count if AGLNDAGRIK2 == .
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.)
```

```
. count if AGLNDAGRIK2 == .
4
```

generate: create a new variable

```
generate agrilog = log10(AGLNDAGRIK2)
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.)
```

```
. generate agrilog = log10(AGLNDAGRIK2)
(4 missing values generated)
```

```
. summarize agrilog
```

Variable		Obs	Mean	Std. Dev.	Min	Max
-----+-----						
agrilog		53	5.579101	.001272	5.576341	5.5804

drop: delete variables

```
drop agrilog
```

```
. use arg_worldbank_2016.dta, clear
(Written by R.          )

. generate agrilog = log10(AGLNDAGRIK2)
(4 missing values generated)

. drop agrilog
```

You can use `keep` if the list of variables you want to retain is shorter than the list of variables you want to delete.

Tougher exercise

Calculate a t-test for a difference in means for the agricultural land between the even and the odd years.

Tip: remember that in math the modulus 2 is 0 for even values and 1 for odd values.

```
. use arg_worldbank_2016.dta
(Written by R.          )

. generate modulus = mod(Year, 2)

. label define moduluslabel 0 "Even" 1 "Odd"

. label values modulus moduluslabel

. ttest AGLNDAGRIK2, by(modulus)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Even	26	379436.2	214.8517	1095.533	378993.7	379878.6
Odd	27	379374.1	220.2221	1144.308	378921.4	379826.7
combined	53	379404.5	152.5084	1110.278	379098.5	379710.6
diff		62.07977	307.9249		-556.1052	680.2647

```
diff = mean(Even) - mean(Odd)          t = 0.2016
Ho: diff = 0                          degrees of freedom = 51
```

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
Pr(T < t) = 0.5795	Pr(T > t) = 0.8410	Pr(T > t) = 0.4205

Some *common* complicated tasks

- Reshaping data sets: wide to long and viceversa
- Merging data sets: Stata ultimately wants you to get all the data you need in a single table
- Looping over a list to do a task
- Working with strings
- Working with dates: there are many different ways to specify time

Finding help

- Try using the user interface then check the code that results from your choices
- Use the `help` command or search the help (top right)
- Google: lots of goodies in older mailing list posts
- Check the UCLA Stata website

More information

- The *Statistics with Stata* by Lawrence Hamilton you already have!
- A quick rundown of the main commands in Stata by the UNC Caroline Population Center
- A recording of an introduction to Stata by Dr Marie Diener-West
- SPSS to Stata table
- UCLA Stata Starter Kit