Introduction to stata

What is stata?

Stata is a powerful statistical software that enables users to analyze, manage, and produce graphical visualizations of data

Why stata?

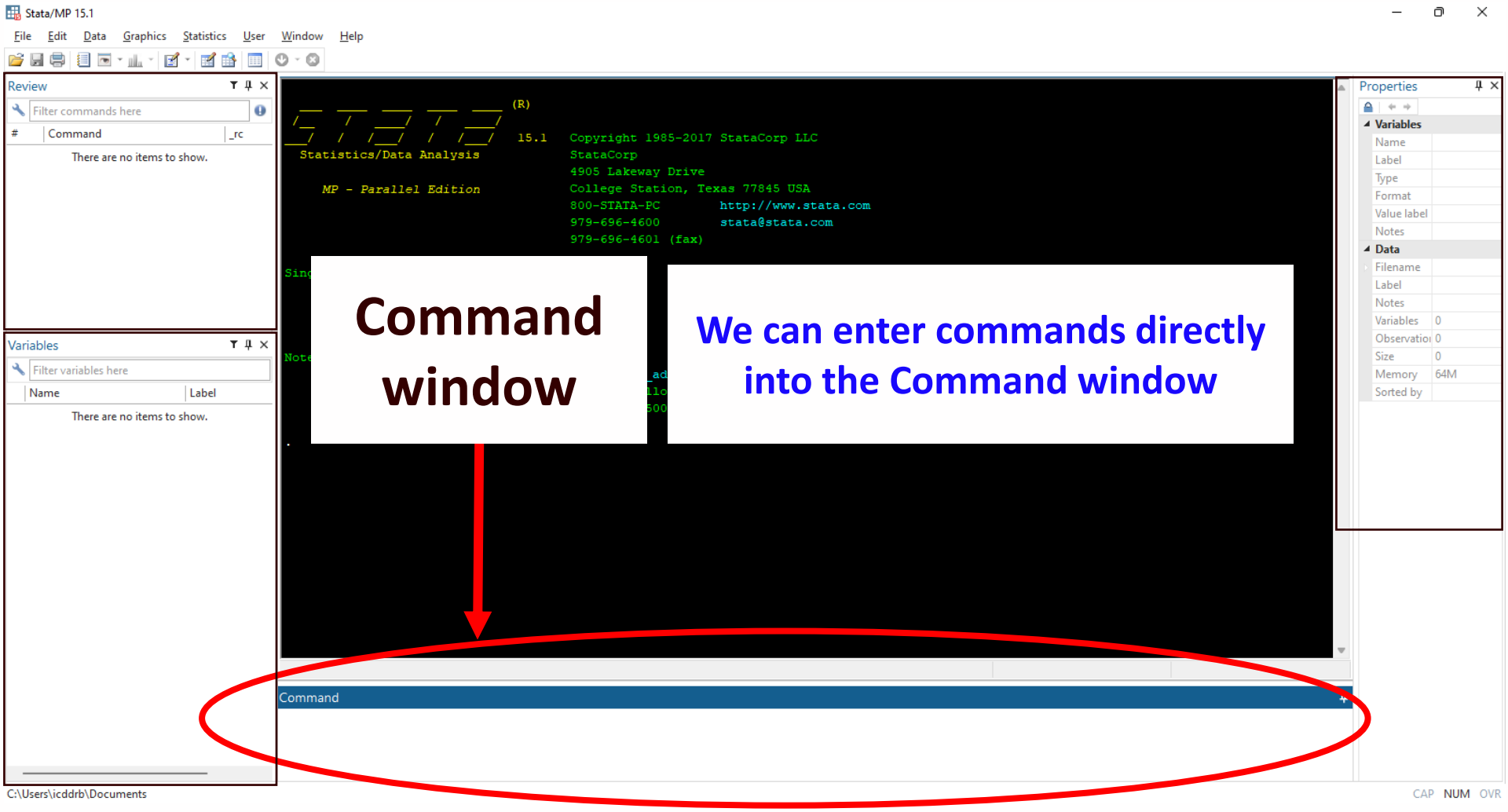
Fast

Accurate

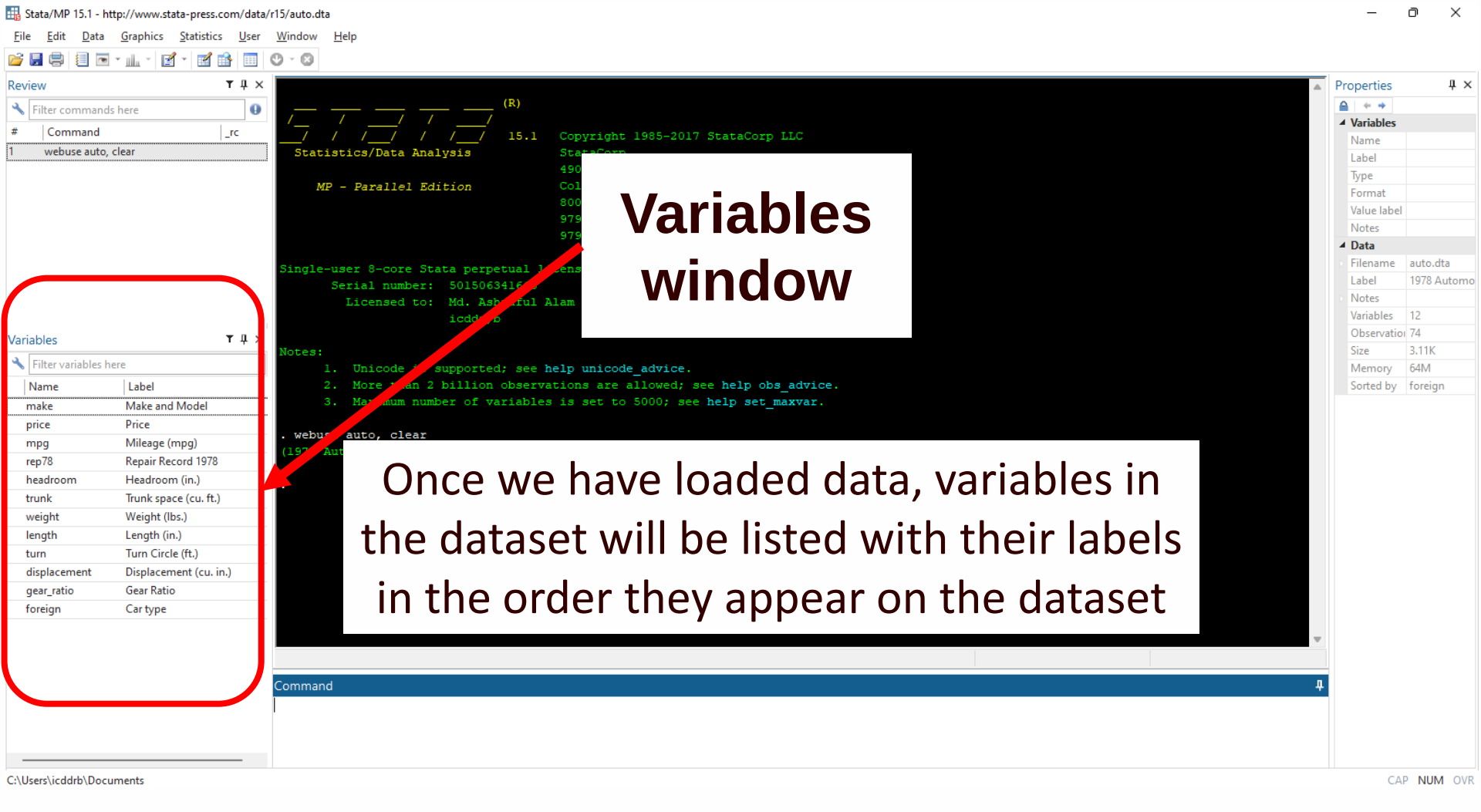
Easy to use

Stata User Interface

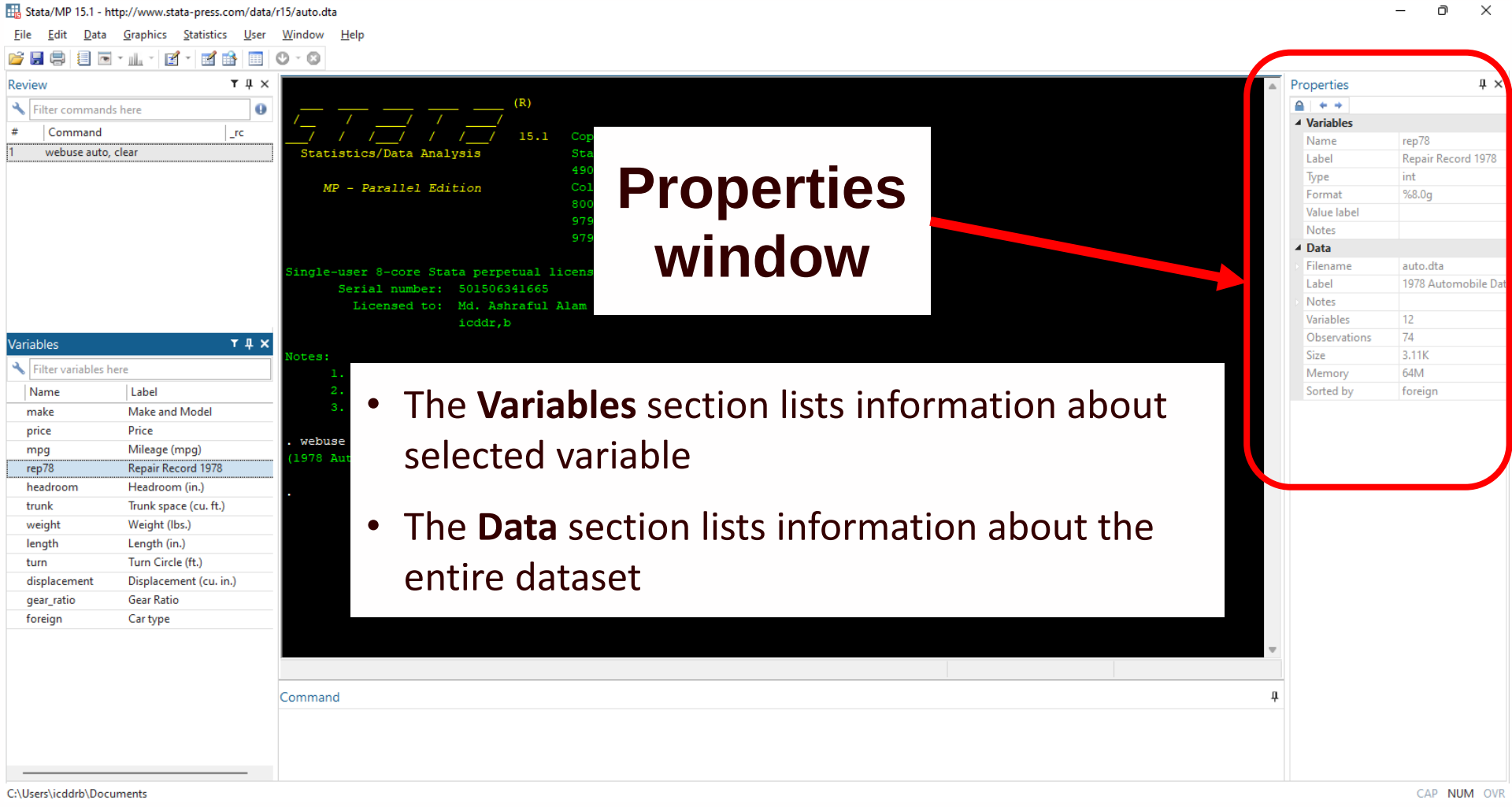
Stata User Interface



Stata User Interface



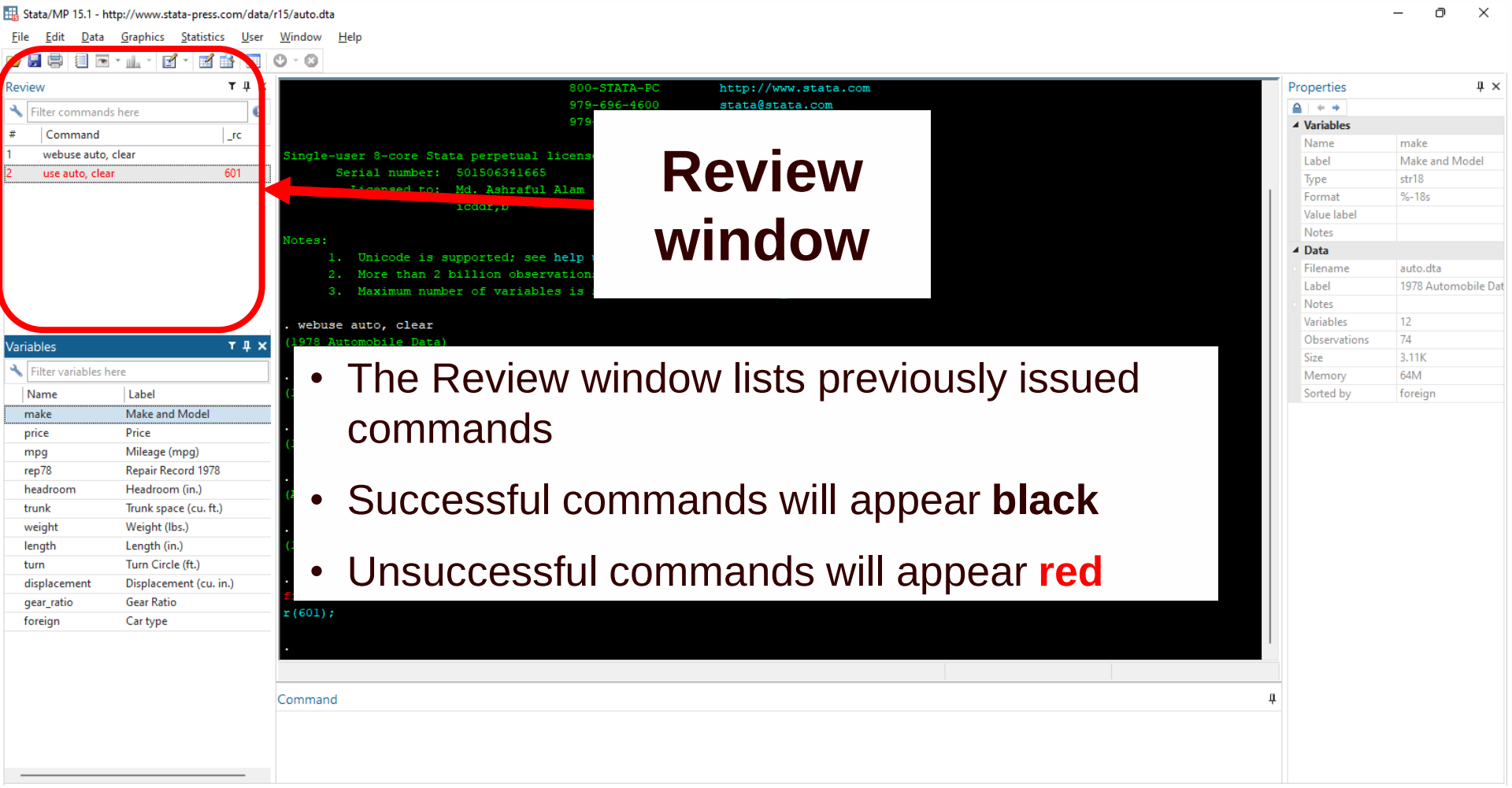
Stata User Interface



Properties window

- The **Variables** section lists information about selected variable
- The **Data** section lists information about the entire dataset

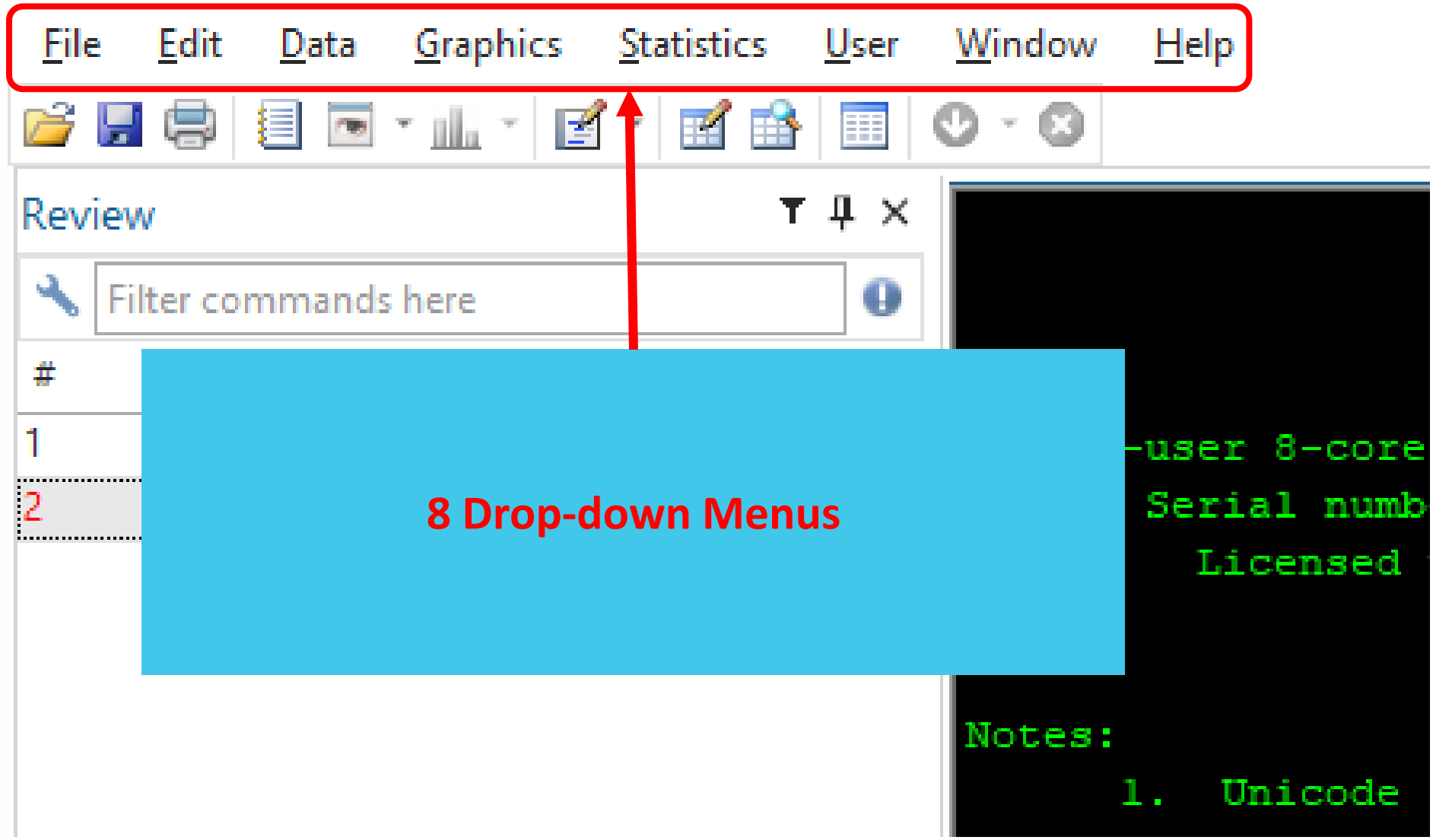
Stata User Interface



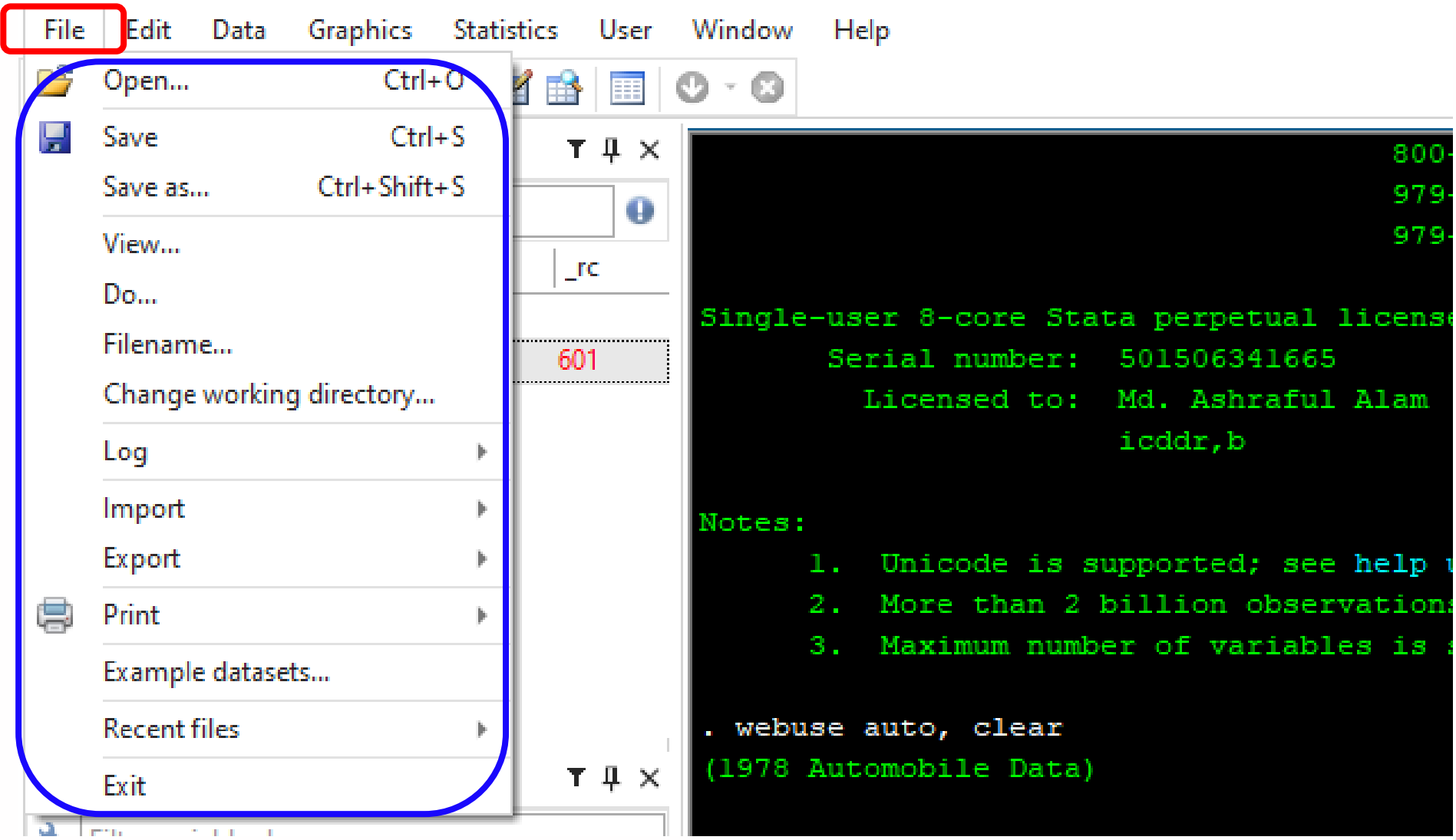
Review window

- The Review window lists previously issued commands
- Successful commands will appear **black**
- Unsuccessful commands will appear **red**

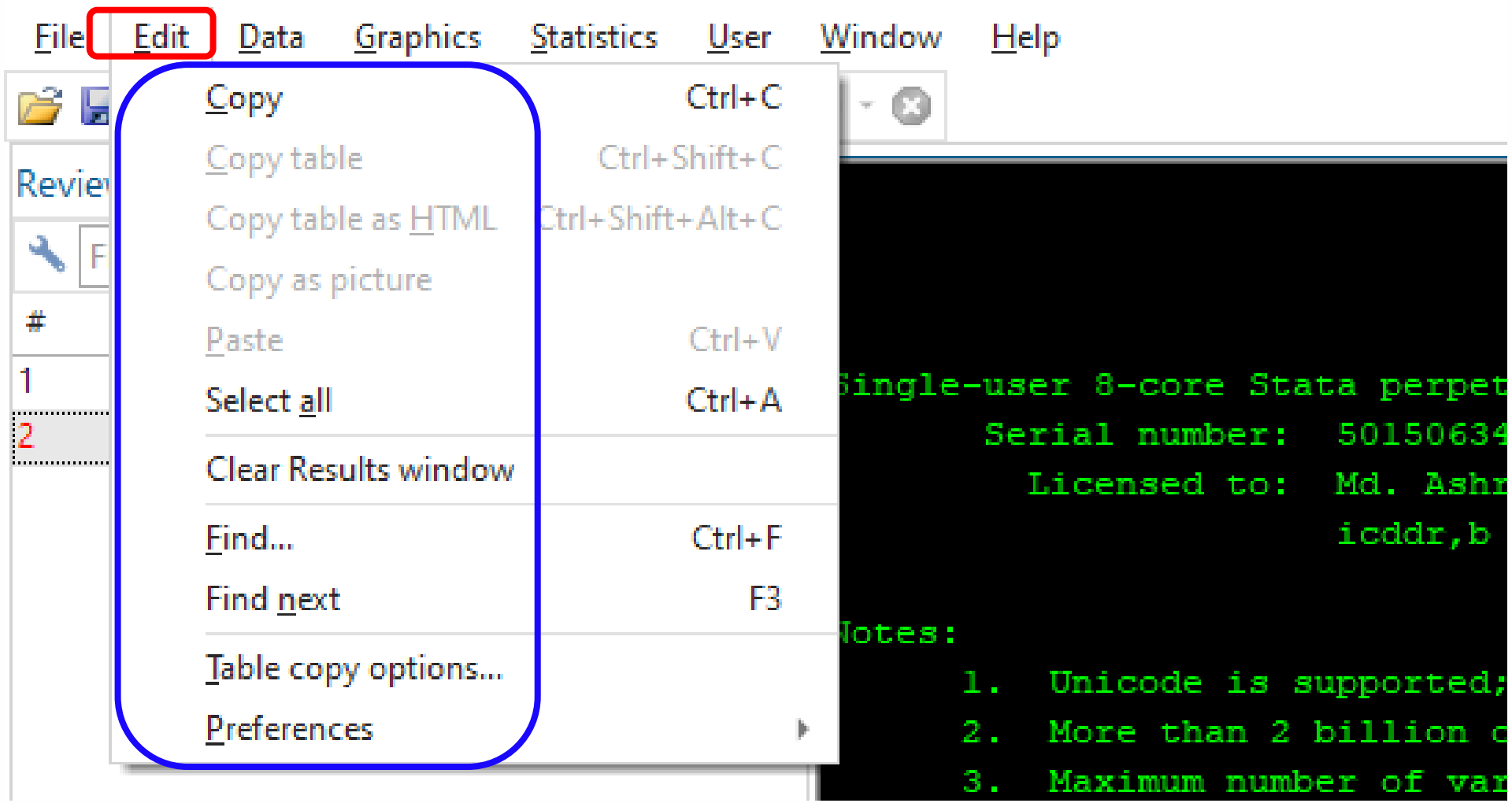
Stata User Interface



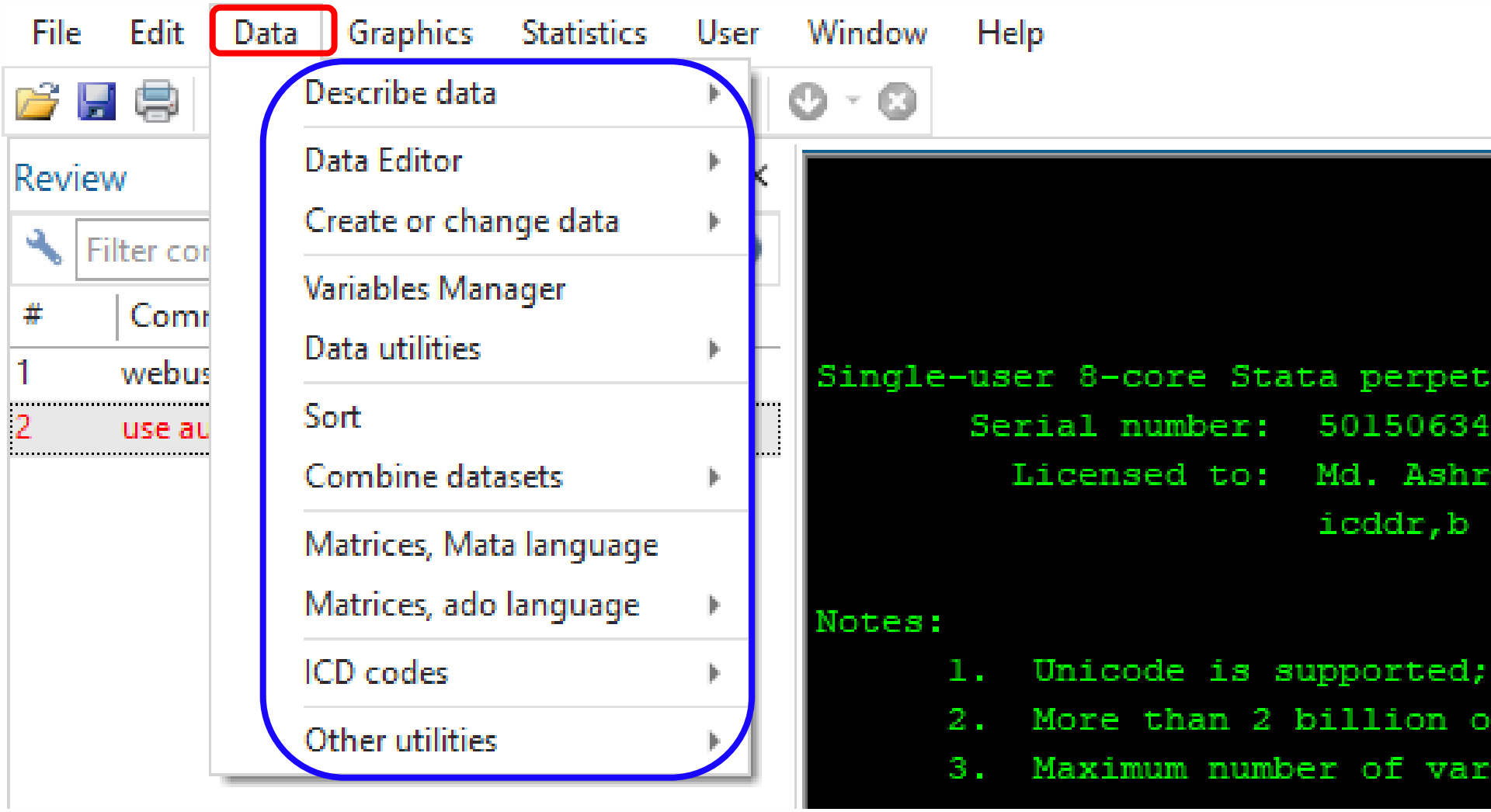
Stata User Interface



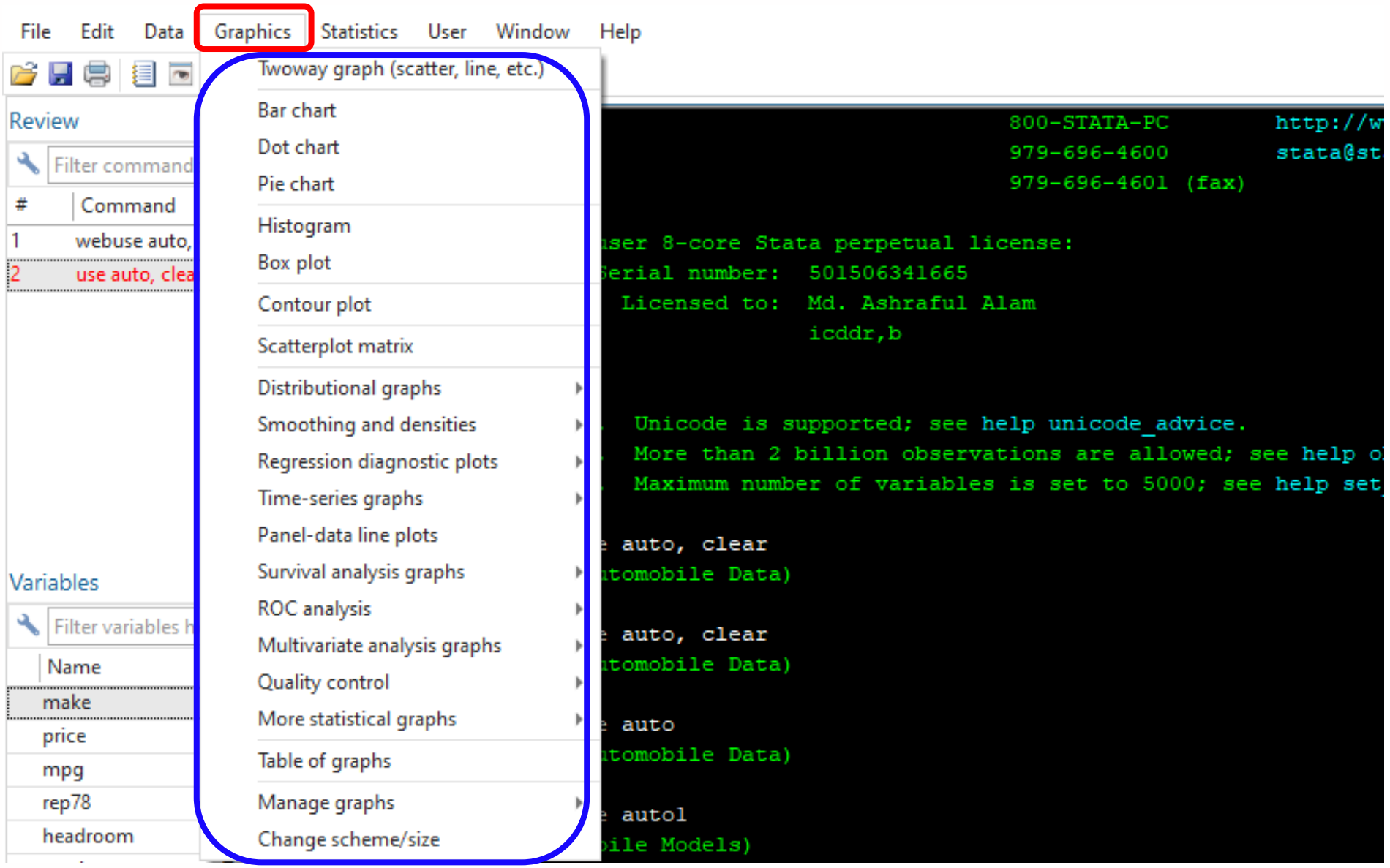
Stata User Interface



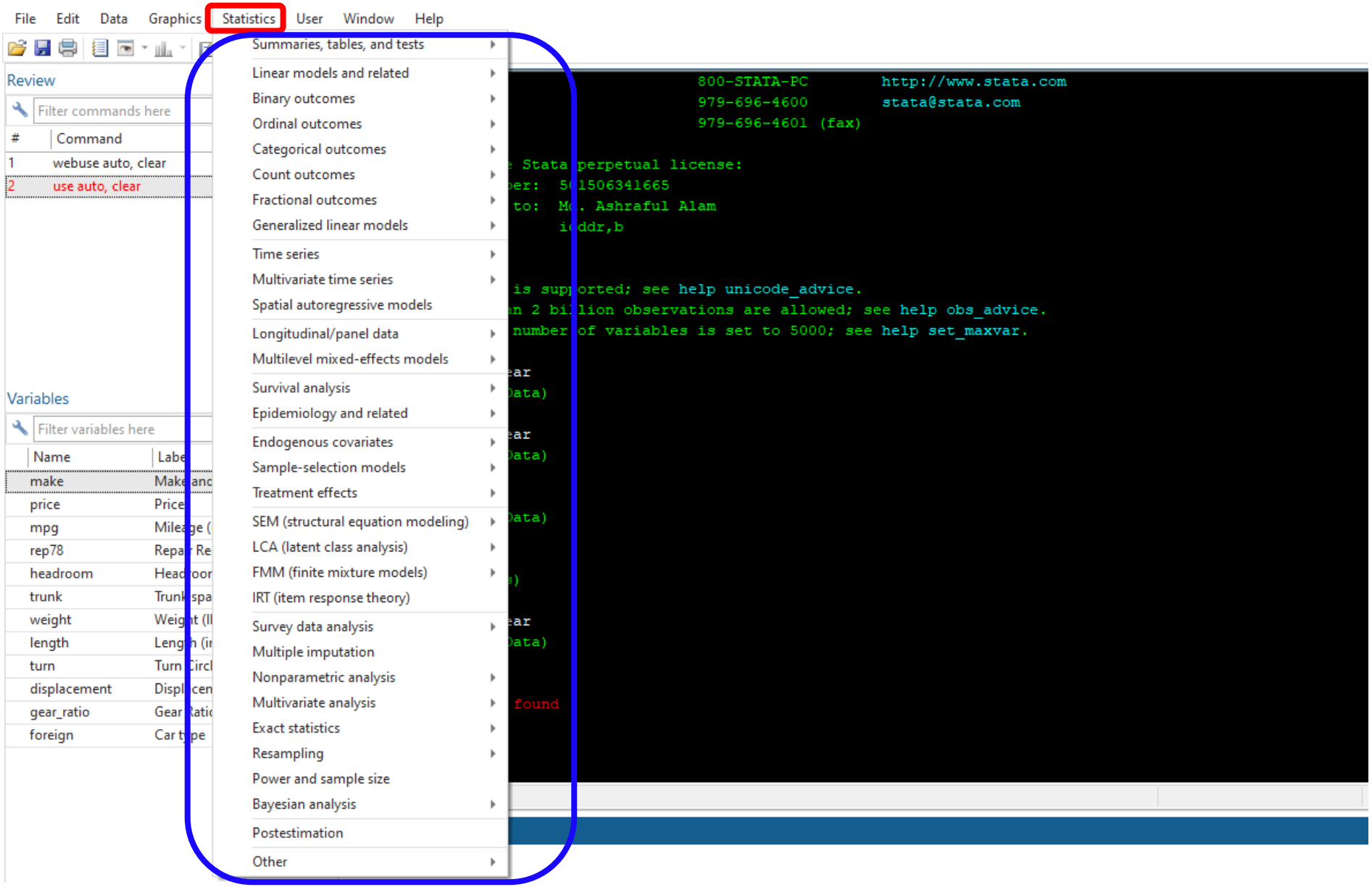
Stata User Interface



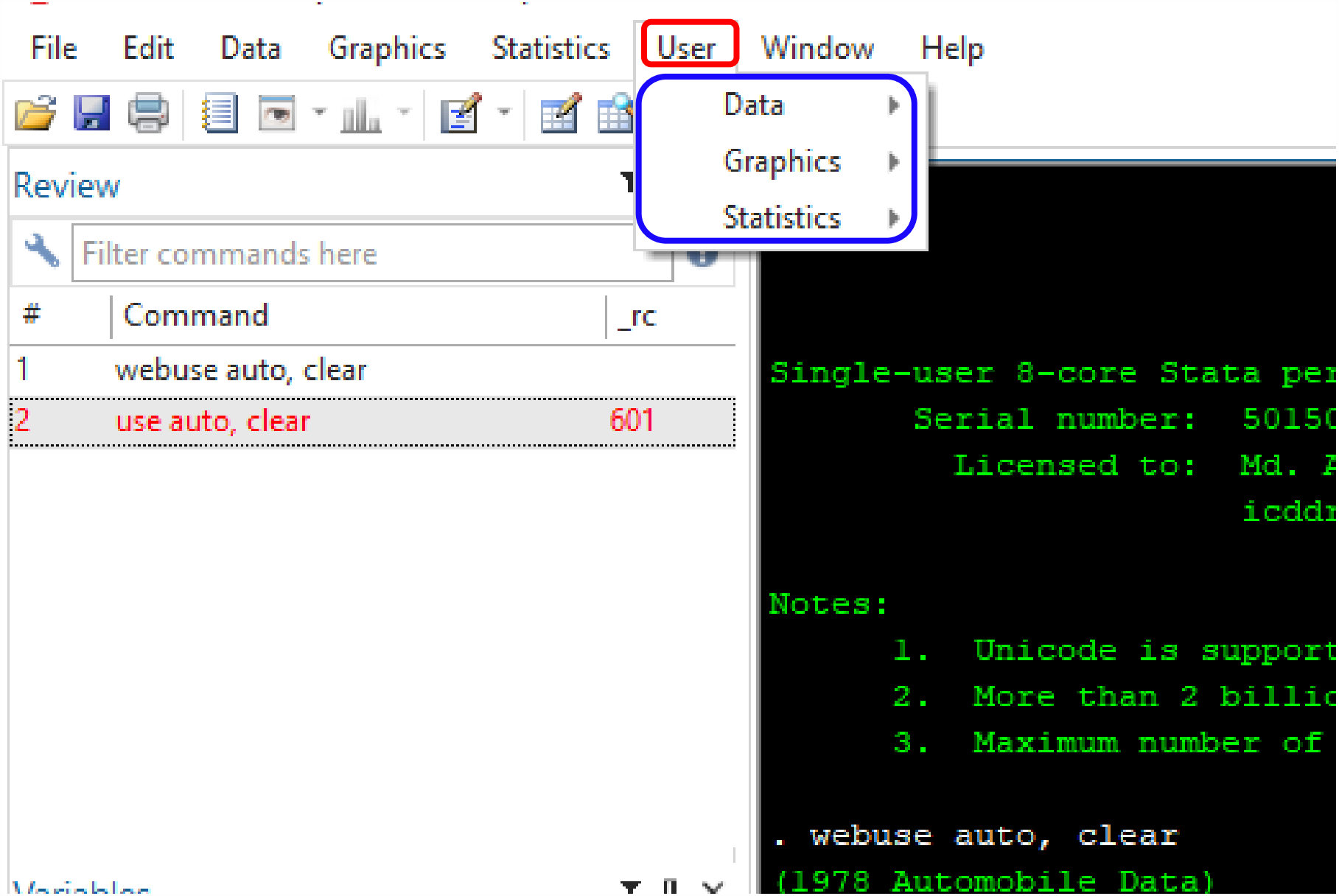
Stata User Interface



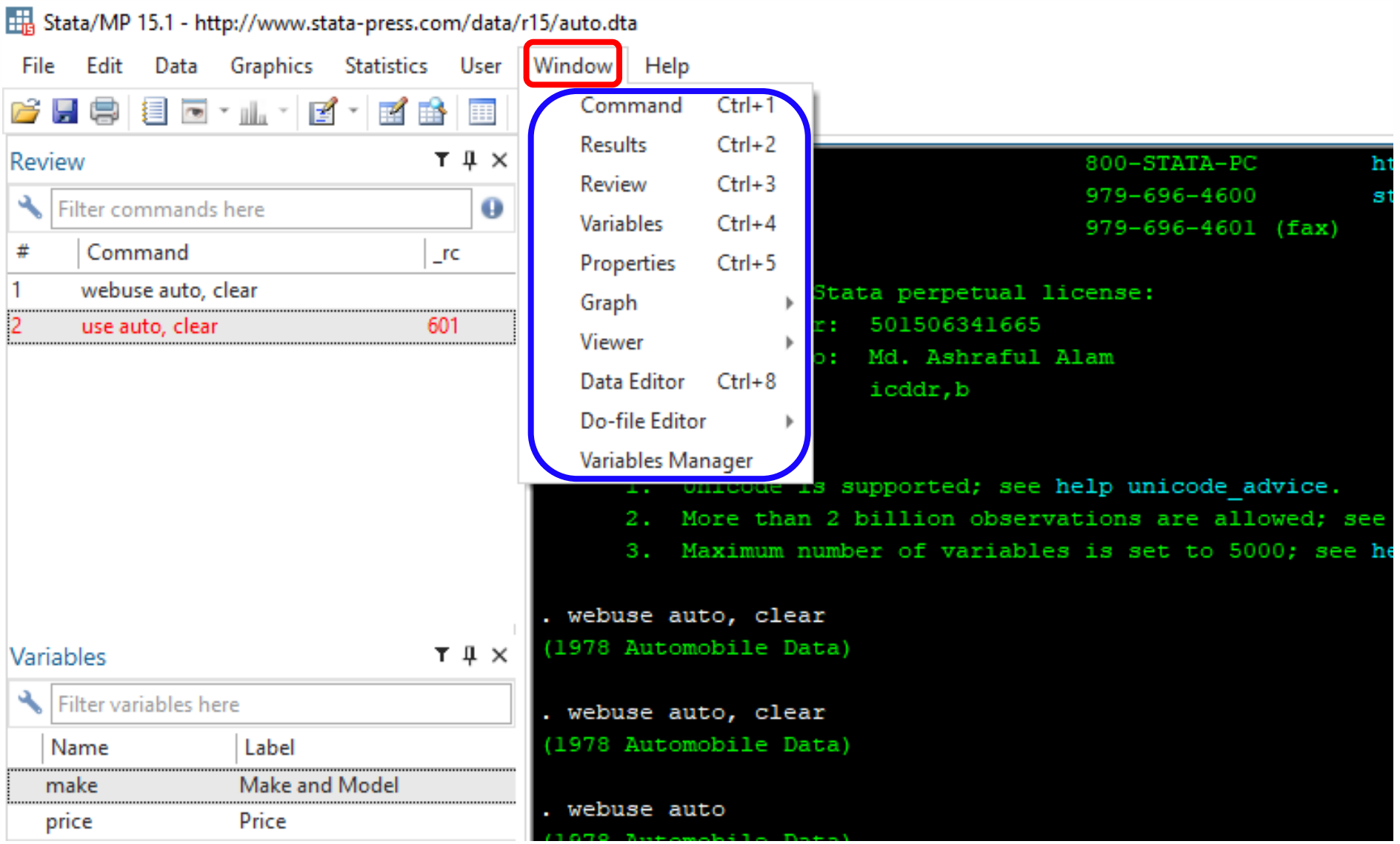
Stata User Interface



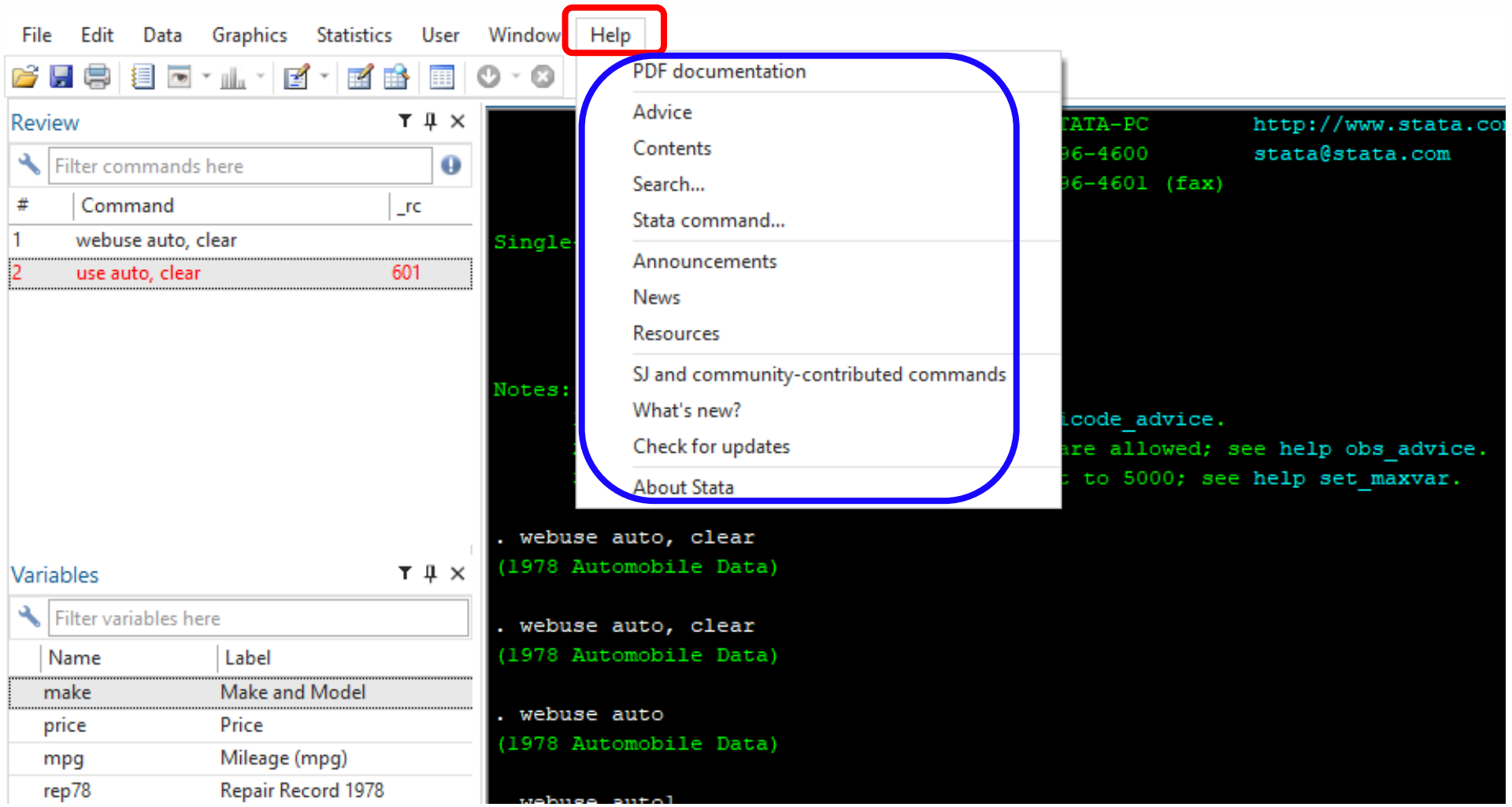
Stata User Interface



Stata User Interface



Stata User Interface



Two Ways of Using Stata

- Interactive mode

  - Using menus and buttons

- Text mode

  –Write command lines in the command window

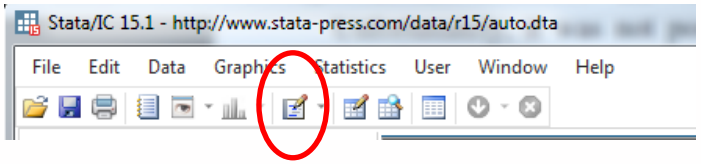
  –Write command lines in a command file (i.e., do file) and execute the do file

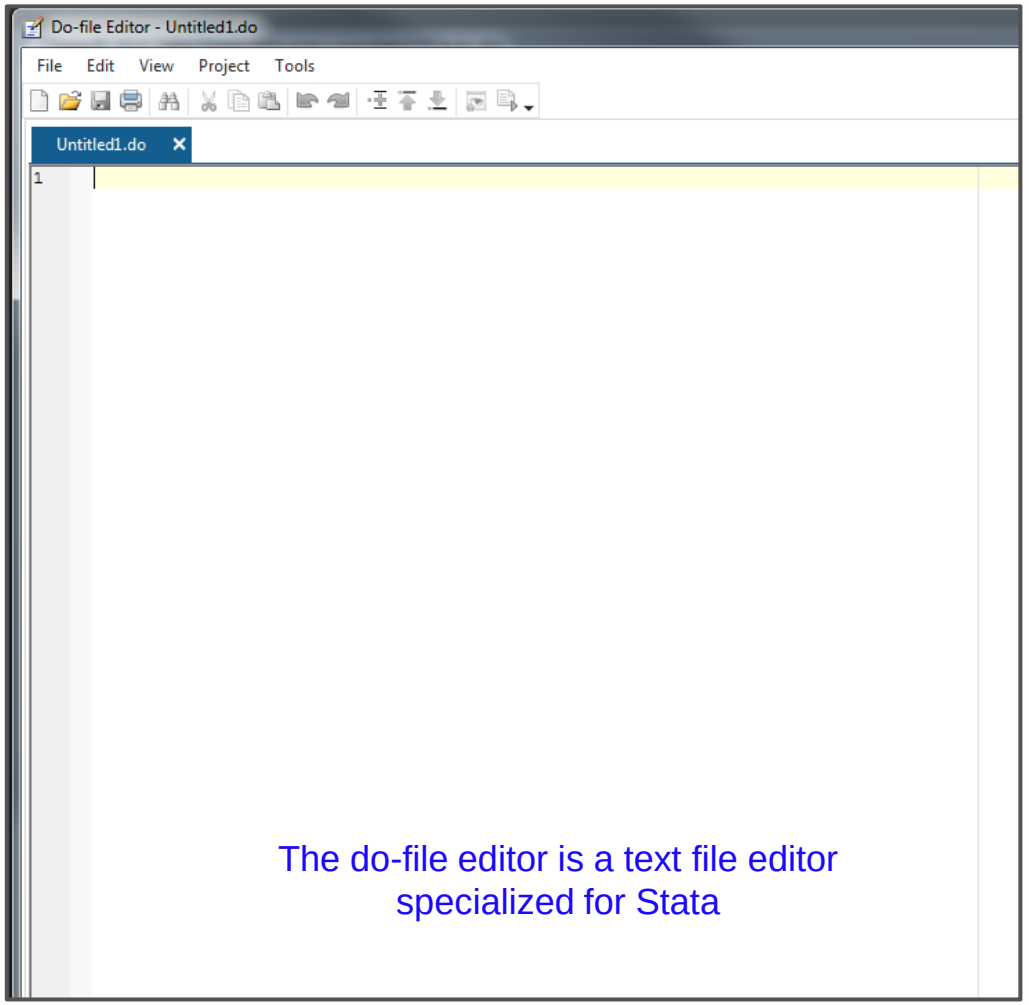
Do-files are scripts of commands

- Stata do-files are text files where users can store and run their commands for reuse, rather than retyping the commands into the Command window
  - ✓Reproducibility
  - ✓Easier debugging and changing commands
- We recommend *always* using a do-file when using Stata

Opening the do-file editor

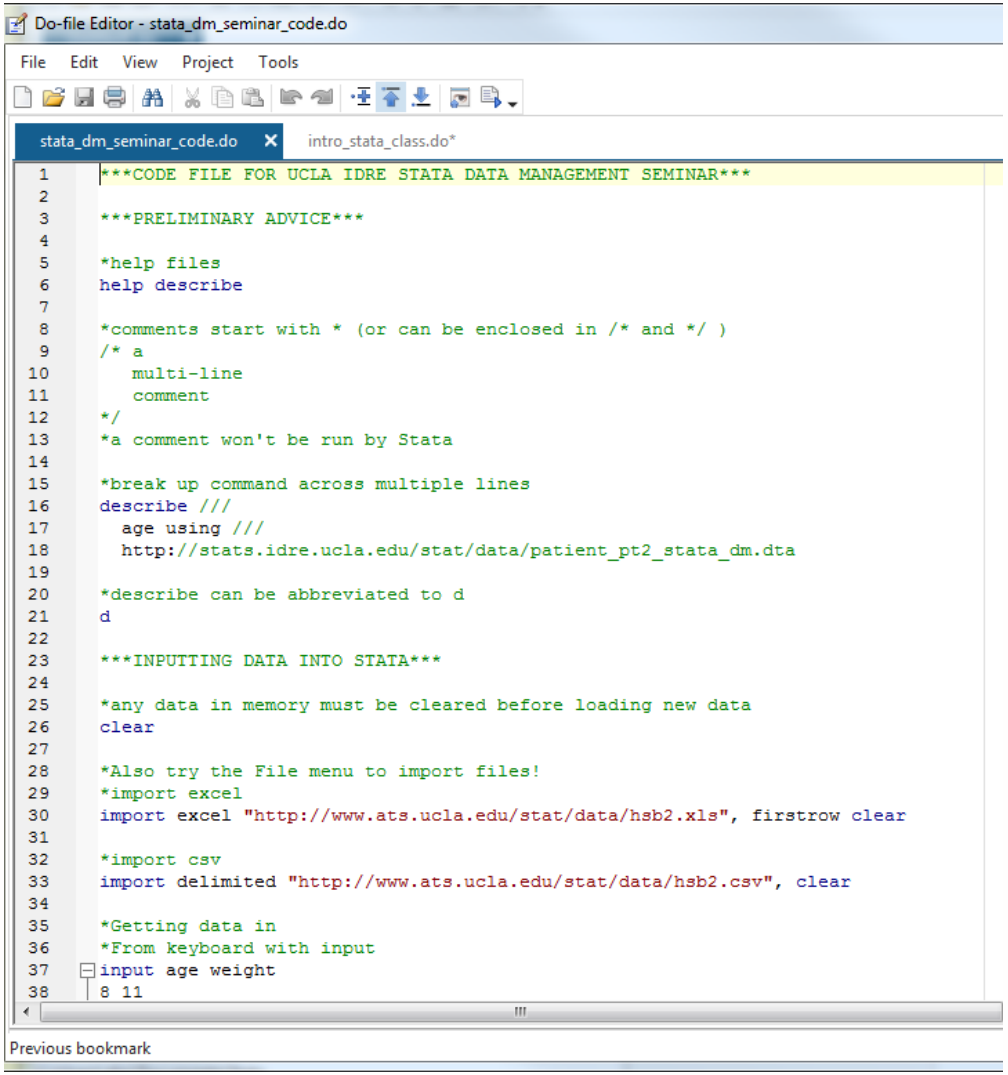
- Use the command **doedit** to open the do-file editor
- Or click on the **pencil and paper** icon on the toolbar





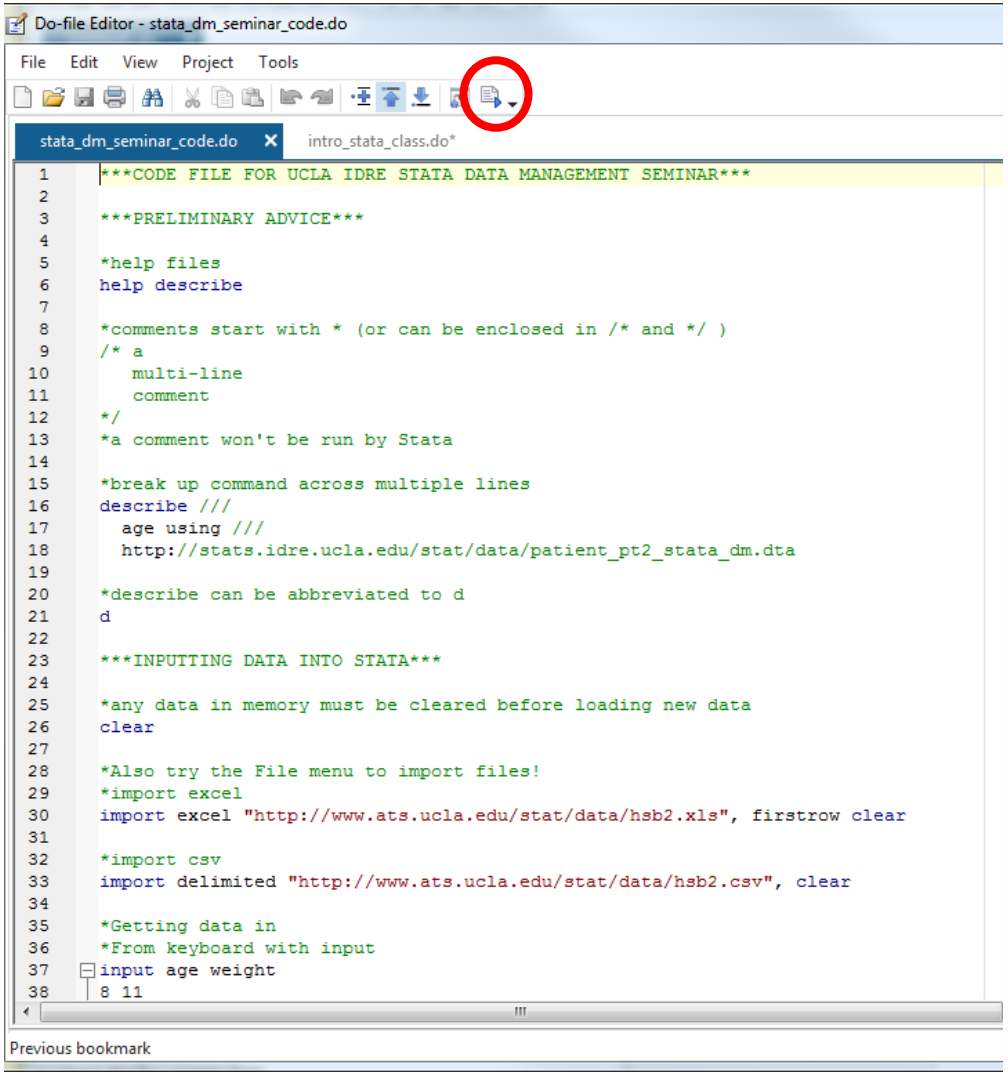
Syntax highlighting

- The do-file editor colors Stata commands **blue**
- Comments, which are not executed, are usually preceded by * and are colored **green**
- Words in quotes (file names, string values) are colored "**red**"



Running commands from the do-file

- To run a command from the do-file, highlight part or all of the command, and then hit **Ctrl+D** or the "**Execute(do)**" icon, the rightmost icon on the do-file editor toolbar
- Multiple commands can be selected and executed



Read Dataset

use	load Stata dataset
save	save Stata dataset
clear	clear dataset from memory
import excel	import Excel dataset
import delimited	import delimited data (csv)

Loading and saving .dta files

- The command **use** loads Stata .dta files
  - Usually these will be stored on a hard drive, but .dta files can also be loaded over the internet (using a web address)
- Use the command **save** to save data in Stata's .dta format
  - The replace option will overwrite an existing file with the same name
- The extension .dta can be omitted when using use and save

```
* read from hard drive; do not execute
use "C:/path/to/myfile.dta"

* load data over internet
use https://stats.idre.ucla.edu/stat/data/hsbdemo

* save data, replace if it exists
save hsbdemo, replace
```

Clearing memory

- Because Stata will only hold one data set in memory at a time, memory must be cleared before new data can be loaded
- The **clear** command removes the dataset from memory
- Data import commands like use will often have a clear option which clears memory before loading the new dataset

```
* load data but clear memory first
use https://stats.idre.ucla.edu/stat/data/hsbdemo, clear
```

Importing excel datasets

- Stata can read in data sets stored in many other formats
- The command **import excel** is used to import Excel data
  - An Excel filename is required (with path, if not located in working directory) after the keyword **using**
- Use the **sheet()** option to open a particular sheet
- Use the **firstrow** option if variable names are on the first row of the Excel sheet

```
* import excel file; change path below before executing
import excel using "C:\path\myfile.xlsx", sheet("mysheet") firstrow clear
```

Importing .csv datasets

- Comma-separated values files are also commonly used to store data
- Use **import delimited** to read in .csv files (and files delimited by other characters such as tab or space)
- The syntax and options are very similar to **import excel**
  - But no need for **sheet()** or **firstrow** options (first row is assumed to be variable names in .csv files)

```
* import csv file; change path below before executing
import delimited using "C:\path\myfile.csv", clear
```

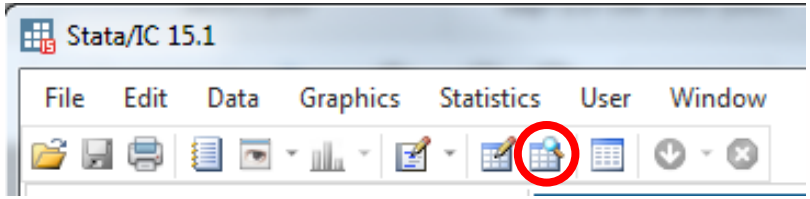
Example dataset

- We will use a dataset consisting of 200 observations (rows) and 11 variables (columns)
- Each observation is a student
- Variables
  - Demographics – gender(1=male, 2=female), race, ses(low, middle, high), etc
  - Academic test scores
    - read, write, math, science, socst

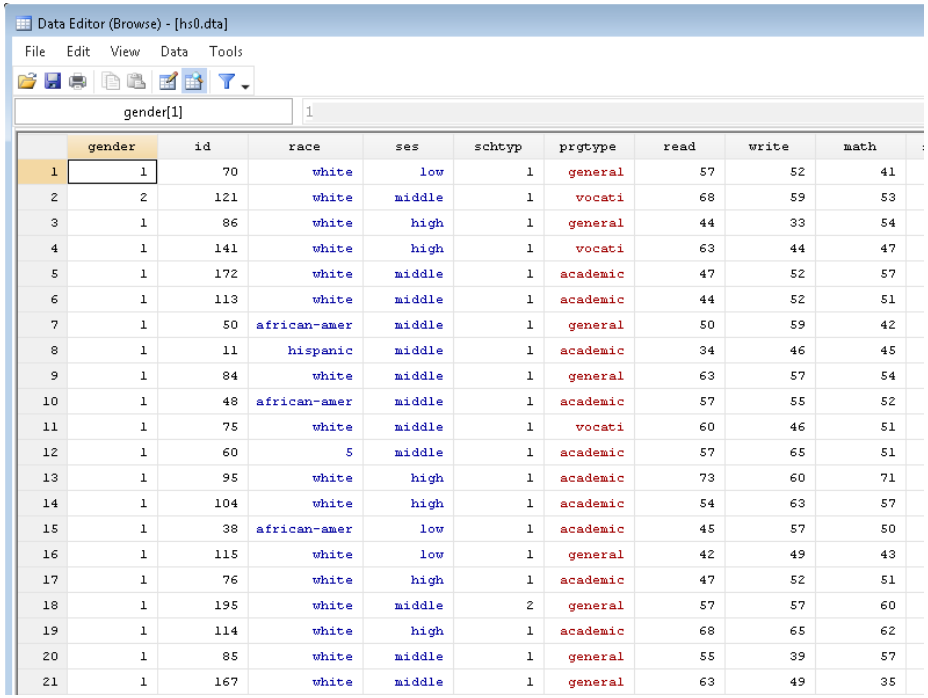
```
* example dataset
use https://stats.idre.ucla.edu/stat/data/hs0, clear
```

Browsing the dataset

- Once the data are loaded, we can view the dataset as a spreadsheet using the command **browse**
- The magnifying glass with spreadsheet icon also browses the dataset



- **Black** columns are numeric, **red** columns are strings, and **blue** columns are numeric with string labels

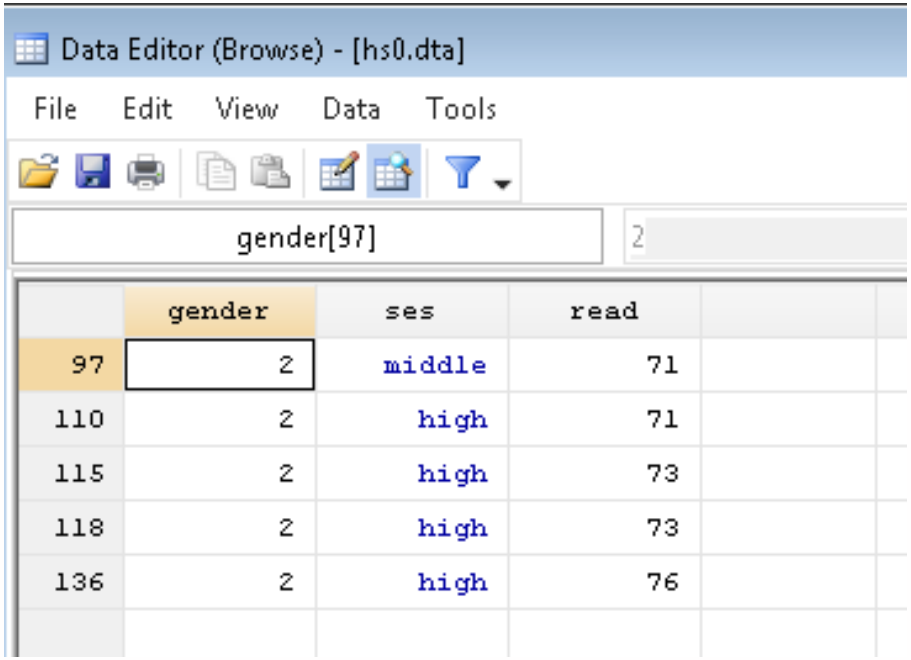


	gender	id	race	ses	schtyp	prgtype	read	write	math
1	1	70	white	low	1	general	57	52	41
2	2	121	white	middle	1	vocati	68	59	53
3	1	86	white	high	1	general	44	33	54
4	1	141	white	high	1	vocati	63	44	47
5	1	172	white	middle	1	academic	47	52	57
6	1	113	white	middle	1	academic	44	52	51
7	1	50	african-amer	middle	1	general	50	59	42
8	1	11	hispanic	middle	1	academic	34	46	45
9	1	84	white	middle	1	general	63	57	54
10	1	48	african-amer	middle	1	academic	57	55	52
11	1	75	white	middle	1	vocati	60	46	51
12	1	60	5	middle	1	academic	57	65	51
13	1	95	white	high	1	academic	73	60	71
14	1	104	white	high	1	academic	54	63	57
15	1	38	african-amer	low	1	academic	45	57	50
16	1	115	white	low	1	general	42	49	43
17	1	76	white	high	1	academic	47	52	51
18	1	195	white	middle	2	general	57	57	60
19	1	114	white	high	1	academic	68	65	62
20	1	85	white	middle	1	general	55	39	57
21	1	167	white	middle	1	general	63	49	35

Stata logical and relational operators

==	equal to double equals used to check for equality
<	less than
>	greater than
<=	less than or equal to
>=	greater than or equal to
!	Not
!=	not equal
&	and
|	or

```
* browse gender, ses, and read
*  for females (gender=2) who have read > 70
browse gender ses read if gender == 2 & read > 70
```



	gender	ses	read
97	2	middle	71
110	2	high	71
115	2	high	73
118	2	high	73
136	2	high	76

Exploring Data

describe	get variable properties
codebook	inspect variable values
summarize	summarize distribution
tabulate	tabulate frequencies

Use describe to get variable properties

- describe provides the following variable properties:
  - storage type (e.g. byte (integer), float (decimal), str8 (character string variable of length 8))
  - name of value label
  - variable label

```
* get variable properties
describe


Contains data from https://stats.idre.ucla.edu/stat/data/hs0.dta
  obs:           200
 vars:            11                          12 Dec 2008 14:38
 size:         9,600
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
gender          float   %9.0g
id              float   %9.0g
race            float   %12.0g     rl
ses             float   %9.0g      sl
schtyp          float   %9.0g
prgtype         str8    %9s
read            float   %9.0g                 reading score
write           float   %9.0g                 writing score
math            float   %9.0g                 math score
science         float   %9.0g                 science score
socst           float   %9.0g                 social studies score
-------------------------------------------------------------------------------
```

Use codebook to inspect variable values

For more detailed information about the values of each variable, use codebook, which provides the following:

- For all variables
  - number of unique and missing values
- For numeric variables
  - range, quantiles, means and standard deviation for continuous variables
  - frequenices for discrete variables
- For string variables
  - frequencies
  - warnings about leading and trailing blanks

```
* inspect values of variables read gender and prgtype
codebook read gender prgtype


-------------------------------------------------------------------------------
read                                                            reading score
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 range:  [28,76]                      units:  1
        unique values:  30                        missing .:  0/200
                  mean:    52.23
             std. dev:   10.2529

           percentiles:        10%       25%       50%       75%       90%
                                39        44        50        60        67

-------------------------------------------------------------------------------
gender                                                            (unlabeled)
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 range:  [1,2]                        units:  1
        unique values:  2                         missing .:  0/200

             tabulation:  Freq.  Value
                            91  1
                           109  2

-------------------------------------------------------------------------------
prgtype                                                           (unlabeled)
-------------------------------------------------------------------------------

                  type:  string (str8)
        unique values:  3                         missing "":  0/200

             tabulation:  Freq.  Value
                           105  "academic"
                            45  "general"
                                50  "vocati"
```

Summarizing continuous variables

- The summarize command calculates a variable's:
  - number of non-missing observations
  - mean
  - standard deviation
  - min and max

```
* summarize continuous variables
summarize read math
```

Variable	Obs	Mean	Std. Dev.	Min	Max
read	200	52.23	10.25294	28	76
math	200	52.645	9.368448	33	75

```
* summarize read and math for females
summarize read math if gender == 2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
read	109	51.73394	10.05783	28	76
math	109	52.3945	9.151015	33	72

Detailed summaries

- Use the detail option with summary to get more estimates that characterize the distribution, such as:
  - percentiles (including the median at 50th percentile)
  - variance
  - skewness
  - kurtosis

```
* detailed summary of read for females
summarize read if gender == 2, detail
```

```
                        reading score
-------------------------------------------------------------
      Percentiles      Smallest
 1%           34            28
 5%           36            34
10%           39            34       Obs                 109
25%           44            35       Sum of Wgt.         109

50%           50                     Mean           51.73394
                      Largest        Std. Dev.      10.05783
75%           57            71
90%           68            73       Variance         101.16
95%           68            73       Skewness       .3234174
99%           73            76       Kurtosis       2.500028
```

Tabulating frequencies of categorical variables

- tabulate displays counts of each value of a variable
  - useful for variables with a limited number of levels
- use the nolabel option to display the underlying numeric values (by removing value labels)

```
* tabulate frequencies of ses
tabulate ses
```

ses	Freq.	Percent	Cum.
low	47	23.50	23.50
middle	95	47.50	71.00
high	58	29.00	100.00
Total	200	100.00	

```
* remove labels
tab ses, nolabel
```

ses	Freq.	Percent	Cum.
1	47	23.50	23.50
2	95	47.50	71.00
3	58	29.00	100.00
Total	200	100.00	

Two-way tabulations

- tabulate can also calculate the joint frequencies of two variables
- Use the row and col options to display row and column percentages

```
* with row percentages
tab race ses, row
```

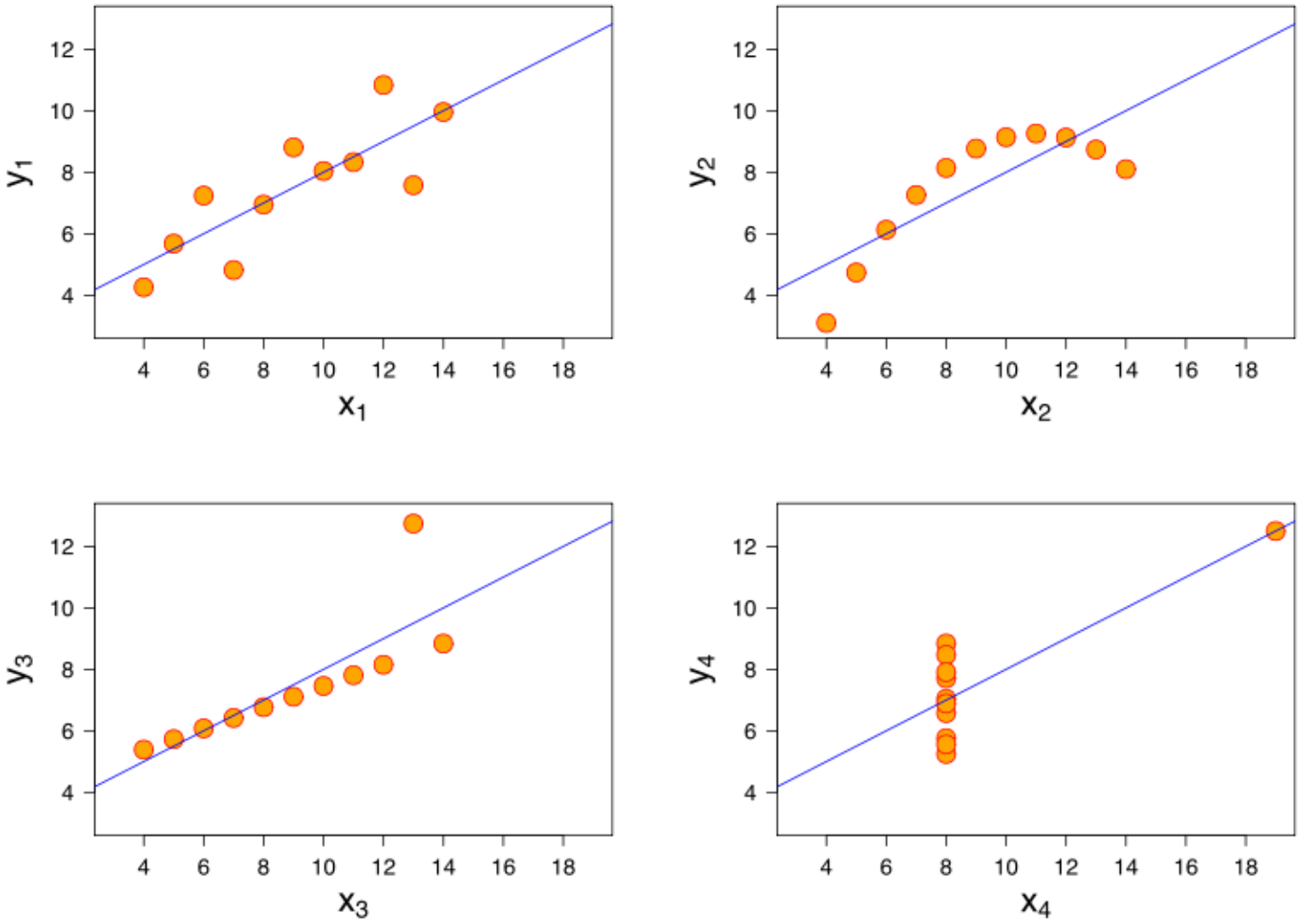
race	ses: low	middle	high	Total
hispanic	9	11	4	24
	37.50	45.83	16.67	100.00
asian	3	5	3	11
	27.27	45.45	27.27	100.00
african-amer	11	6	3	20
	55.00	30.00	15.00	100.00
white	24	71	48	143
	16.78	49.65	33.57	100.00
5	0	2	0	2
	0.00	100.00	0.00	100.00
Total	47	95	58	200
	23.50	47.50	29.00	100.00

Importance of data visualization

Anscombe's quartet datasets:

Property	Value
Mean of *x* in each case	9 (exact)
Sample variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Sample variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal ….places, respectively)

Importance of data visualization



icddr,b

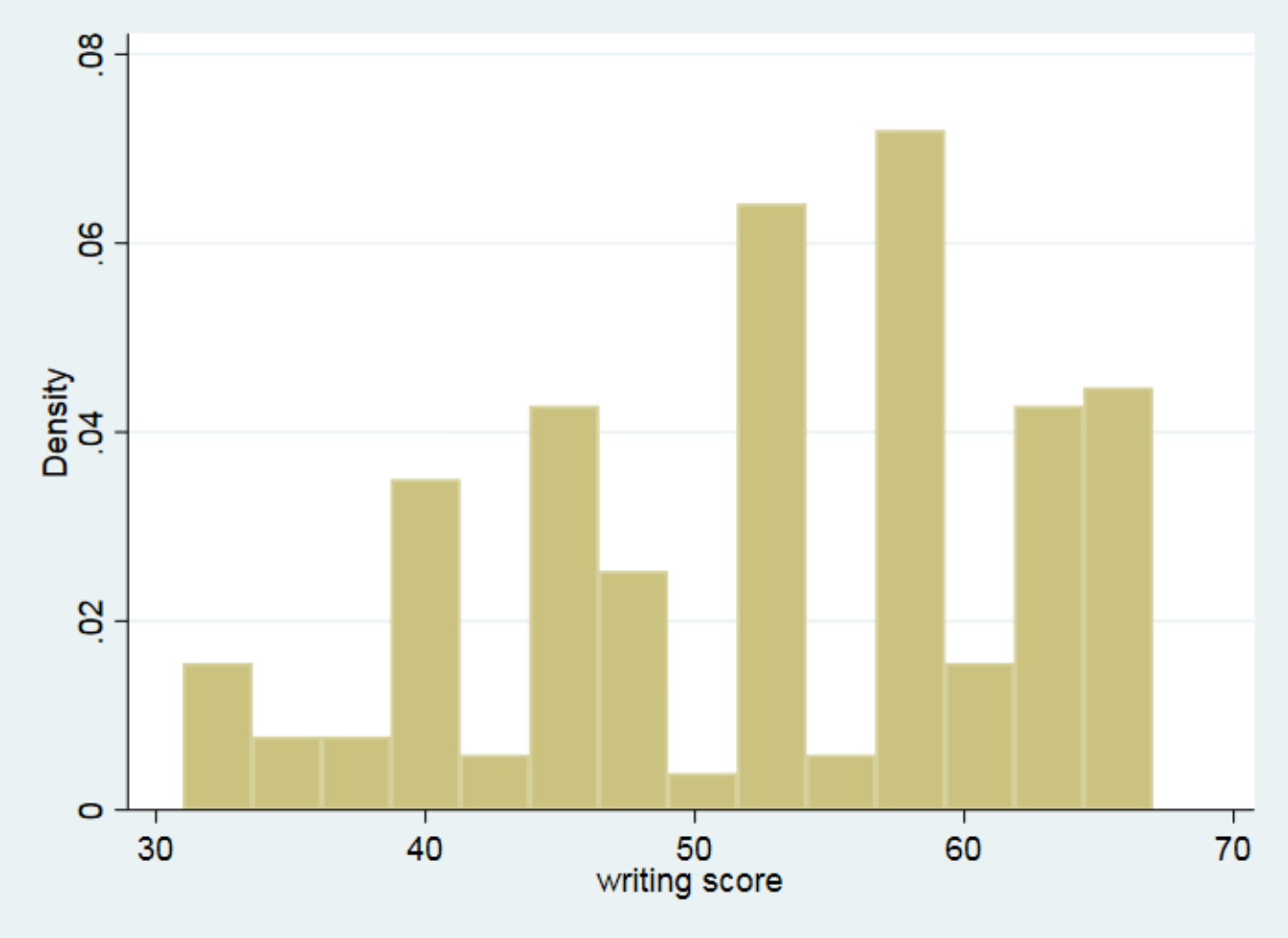
Importance of data visualization

Anscombe's quartet

I		II		III		IV	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

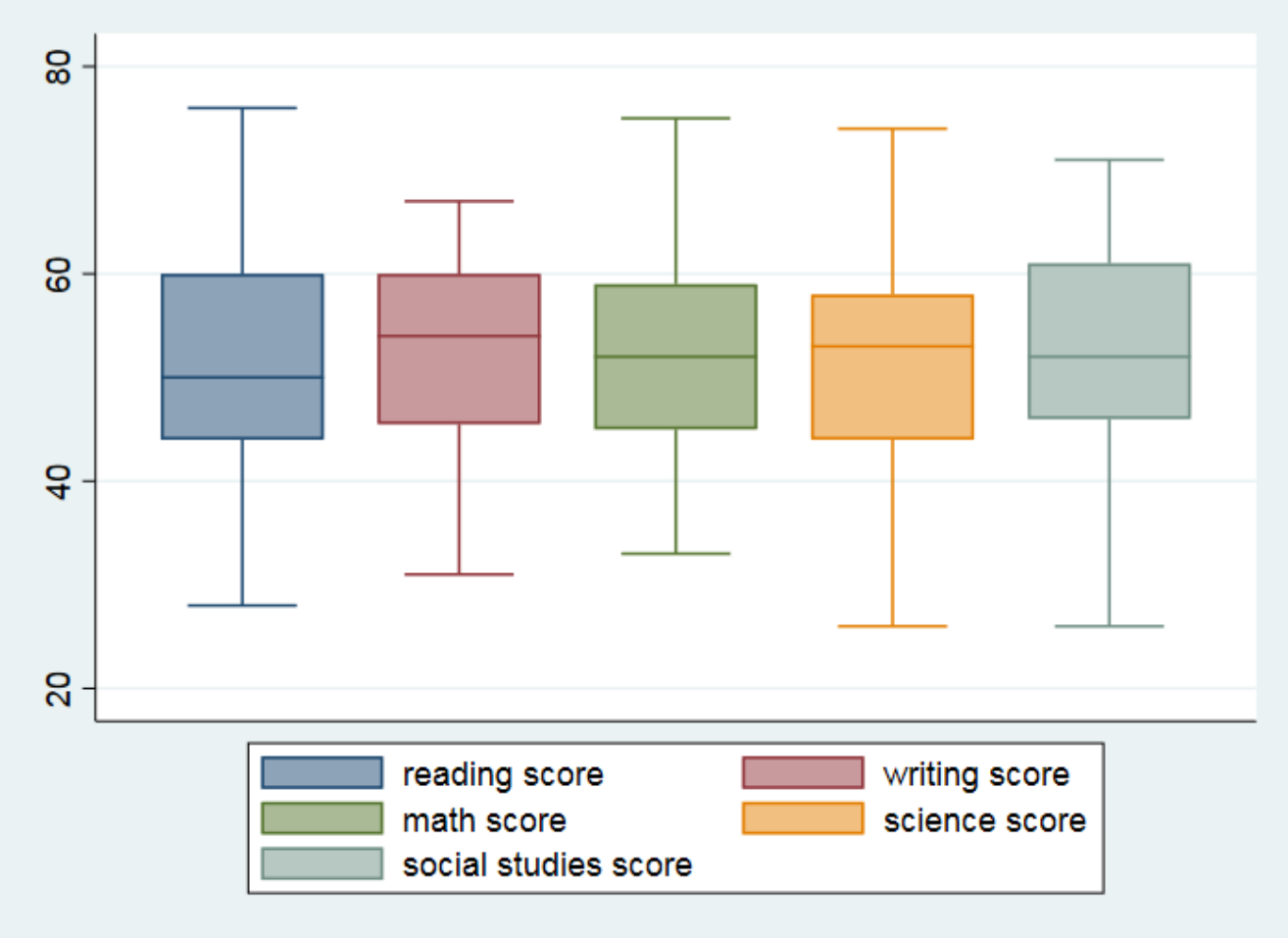
Histograms

```
*histogram of write
histogram write
```



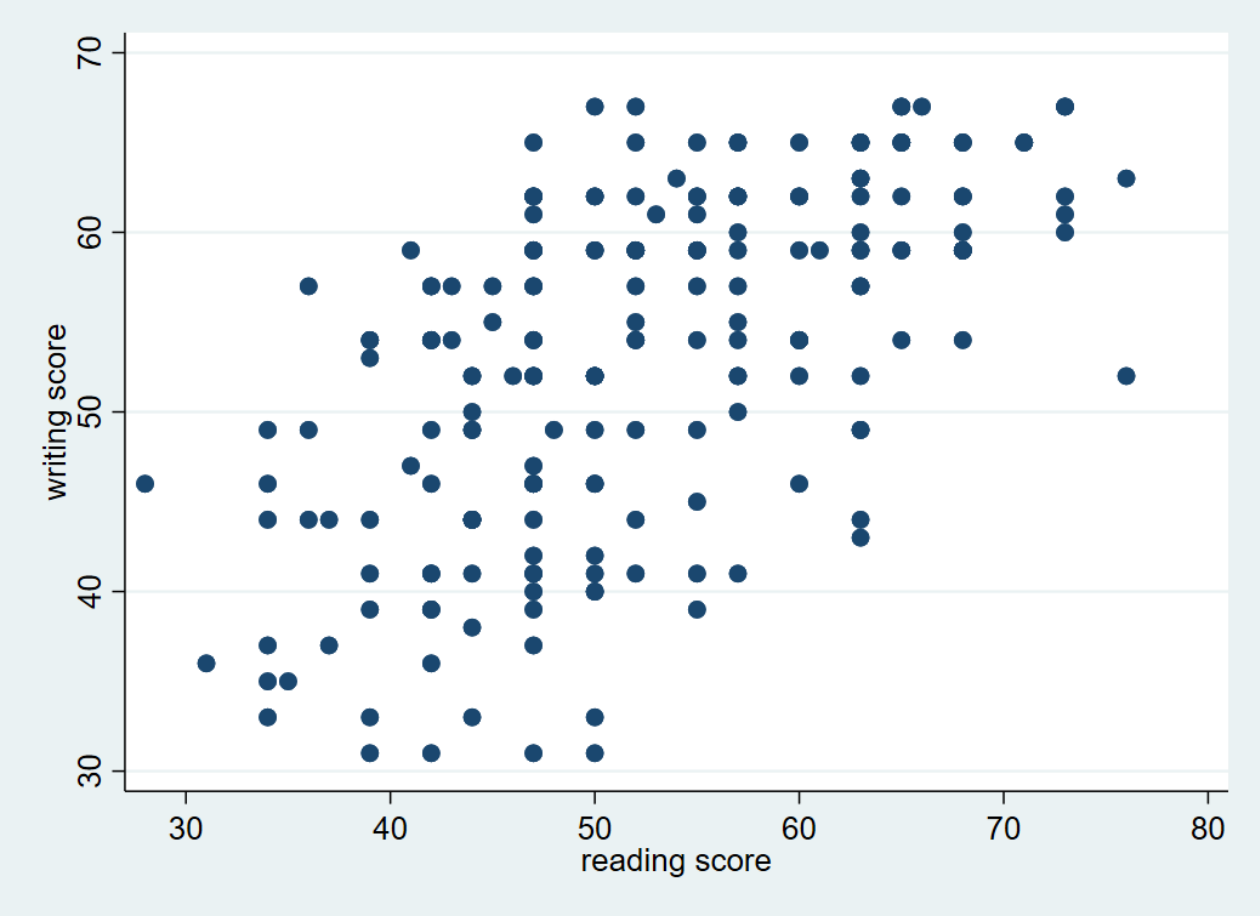
Boxplots

```
* boxplot of all test scores
graph box read write math science socst
```



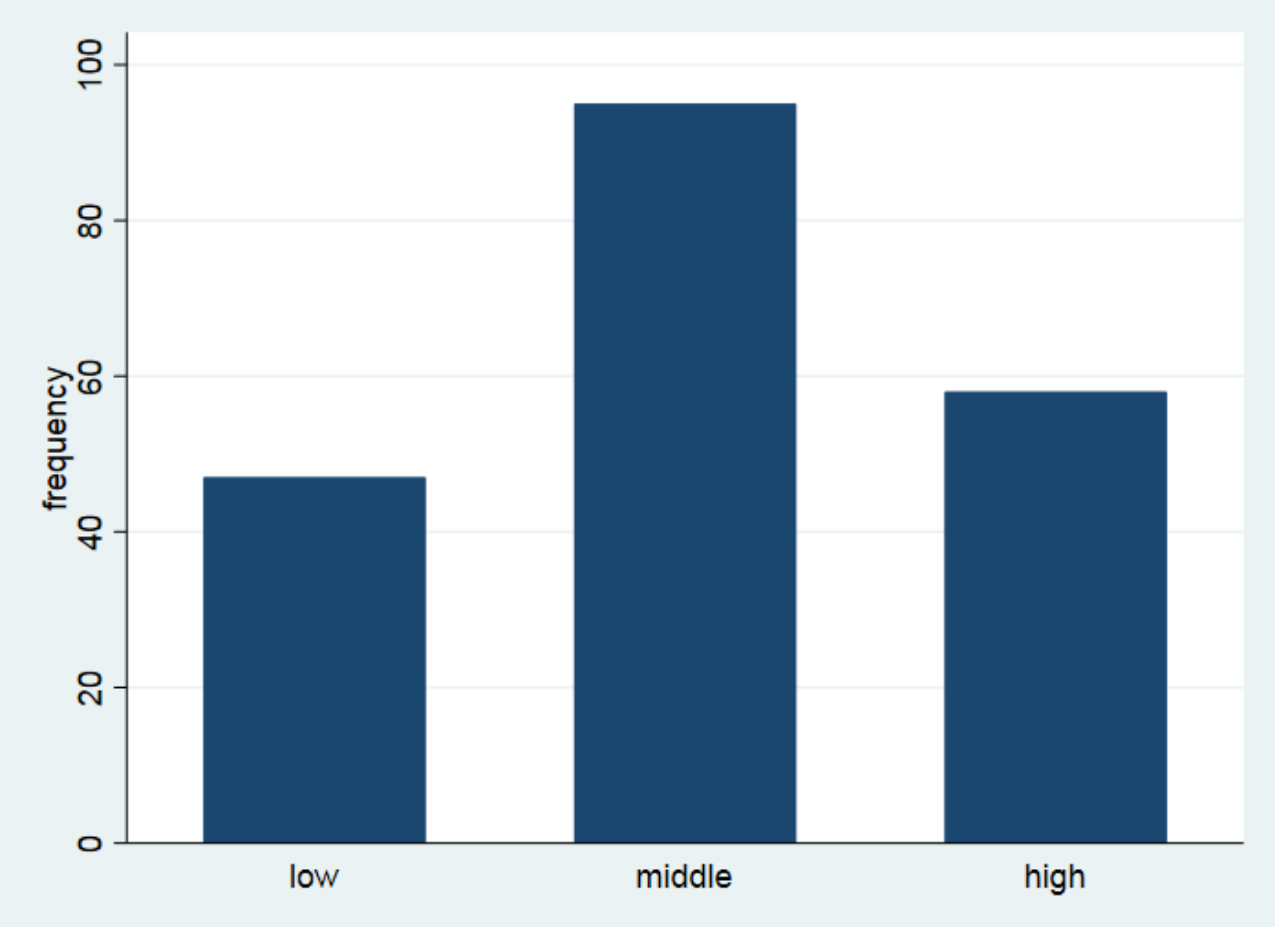
Scatter plots

```
* scatter plot of write vs read
scatter write read
```



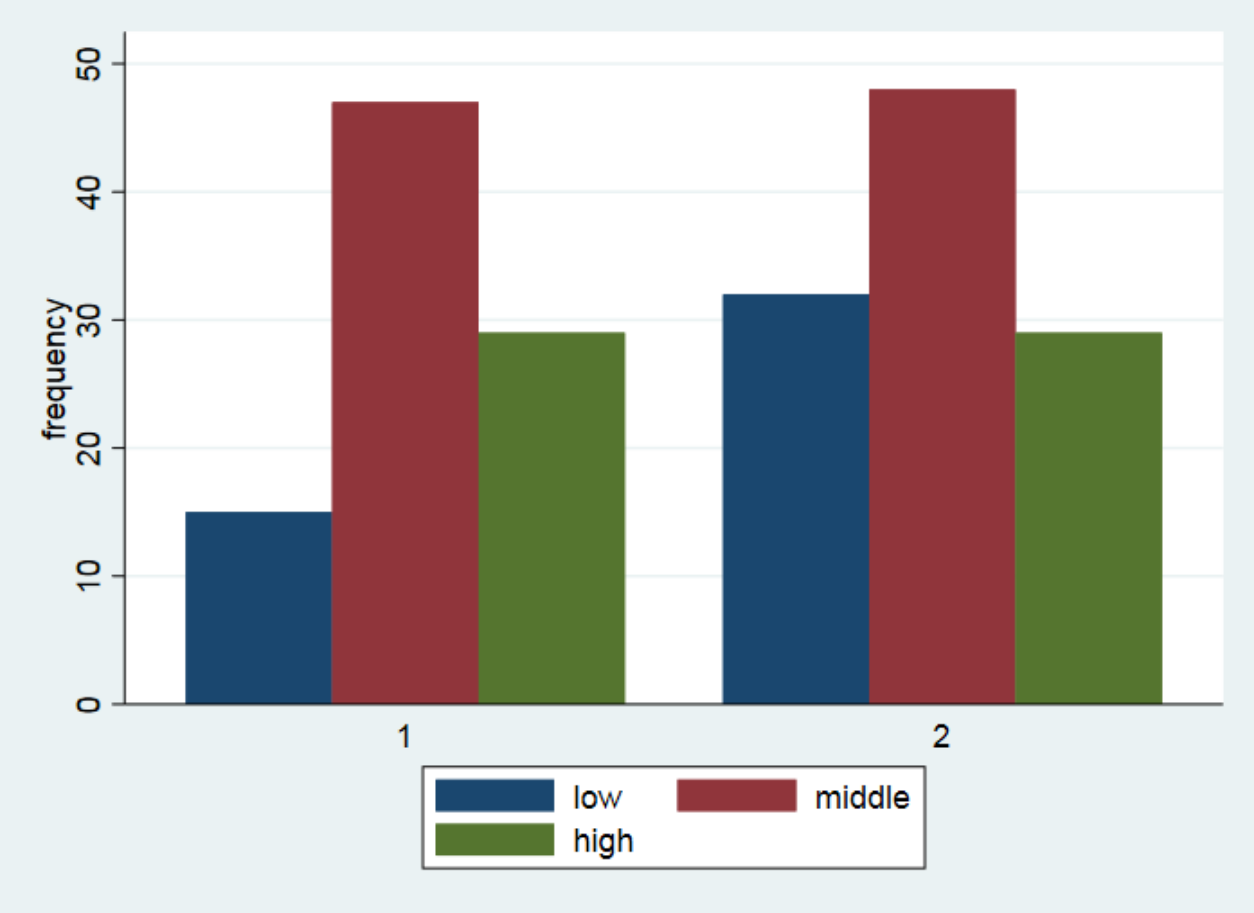
Bar graphs to visualize frequencies

```
* bar graph of count of ses
graph bar (count), over(ses)
```



Two-way bar graphs

```
* frequencies of gender by ses
*   asyvars colors bars by ses
graph bar (count), over(ses) over(gender) asyvars
```



Basic Statistical Analysis

Means and confidence intervals

```
* many commands provide 95% CI
mean read

Mean estimation                   Number of obs    =       200

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        read |      52.23   .7249921       50.80035    53.65965
--------------------------------------------------------------
```

Means and confidence intervals

We can change the confidence level of the interval with the ci means command and the **level()** option

```
* 99% CI for read
ci means read, level(99)
```

```
    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
        read |        200       52.23    .7249921        50.34447    54.11553
```

Independent samples t-test example

```
* independent samples t-test
ttest read, by(female)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      91    52.82418    1.101403    10.50671    50.63605     55.0123
       1 |     109    51.73394    .9633659    10.05783    49.82439     53.6435
---------+--------------------------------------------------------------------
combined |     200       52.23    .7249921    10.25294    50.80035    53.65965
---------+--------------------------------------------------------------------
    diff |            1.090231    1.457507               -1.783998    3.964459
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.7480
Ho: diff = 0                     degrees of freedom =      198

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7723         Pr(|T| > |t|) = 0.4553          Pr(T > t) = 0.2277
```

Correlation

- A correlation coefficient quantifies the linear relationship between two (continuous) variables on a scale between -1 and 1
- The output will be a correlation matrix that shows the pairwise correlation between each pair of variables

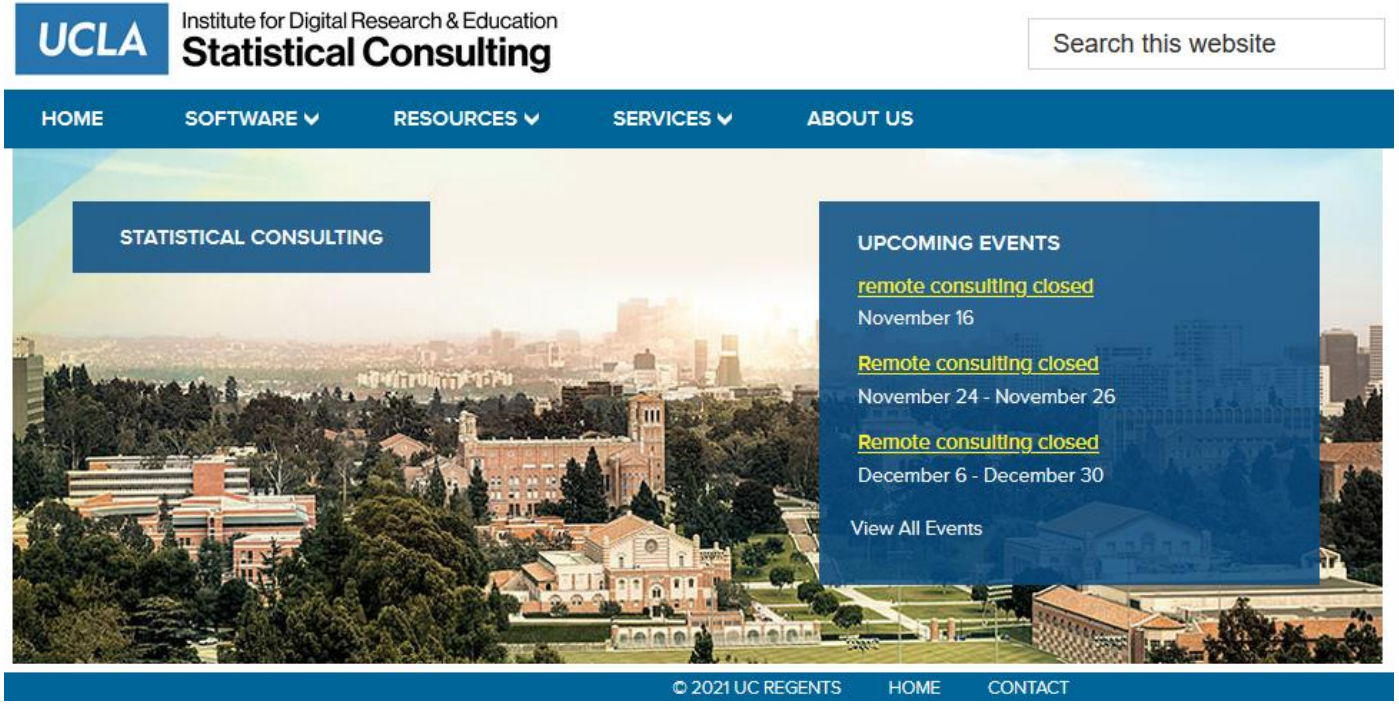
```
* correlation matrix of 5 variables
corr read write math science socst

(obs=195)

             |     read    write     math  science     socst
-------------+------------------------------------------------
        read |   1.0000
       write |   0.5960   1.0000
        math |   0.6492   0.6203   1.0000
     science |   0.6171   0.5671   0.6166   1.0000
       socst |   0.6175   0.5996   0.5299   0.4529   1.0000
-------------------------------------------------------------
```

IDRE statistical consulting website

- The IDRE Statistical Consulting website is a well-known resource for coding support for several statistical software packages
  - https://stats.idre.ucla.edu
- Stata was beloved by previous members of the group, so Stata is particularly well represented on our website



Thank you

icddr,b thanks its core donors for their on-going support

Government of the People's Republic of Bangladesh

Canada

SWEDEN

UKaid from the British people