



STATA TRAINING

Shaheed Bhagat Singh College

Shweta Gupta

Research Analyst

Environment & Production Technology Division (EPTD)

International Food Policy Research Institute

New Delhi | 1st April 2022

		Topics covered
DAY 1	Part 1	Introduction to STATA & its components
	Part 2	Understanding data
	Part 3	Data transformation
	Part 4	Data visualization
DAY 2	Part 5	Data cleaning
	Part 6.1	Regression analysis
	Part 6.2	Different functional forms
	Part 6.3	Exploring CLRM assumptions
		Assignment
DAY 3	Part 7	Types of data
	Part 8	Monte Carlo Experiment
		Discussing Assignment

Part 1: Introduction to STATA & its components



Review

T ↕ ×

Filter commands ⓘ

Command | _rc

There are no items to show.

```
----- (R)
-----
Statistics/Data Analysis 15.0 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)

Single-user Stata perpetual license:
Serial number: 301506215585
Licensed to: www.Downloadly.ir
Iran Will Defeat US

Notes:
1. Unicode is supported; see help unicode_advice.
```

Command

Variables

T ↕ ×

Filter variables here

Name | Label

There are no items to show.

Properties

↕ ×

🔒 ⬅ ➡

Variables

Name

Label

Type

Format

Value label

Notes

Data

Filename

Label

Notes

Toolbar

**All
previously
executed
commands
show here**

Filter commands here

#	Command
1	sysuse auto.dta
2	summarize

15.0 Copyright 1985-2017 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)

Single-user Stata perpetual license:
 Serial number: 301506215585
 Licensed to: www.Downloadly.ir
 Iran Will Defeat US

Notes:
 1. Unicode is supported; see [help unicode_advice](#).

```
. sysuse auto.dta
(1978 Automobile Data)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

Command

Version number

**Results from
executing
commands**

Type command here to execute

Filter variables here

Name	Label
make	Make and Model
price	Price
mpg	Mileage (mpg)
rep78	Repair Record 1978
headroom	Headroom (in.)
trunk	Trunk space (cu. ft.)
weight	Weight (lb.)
length	Length (in.)
turn	Turn (degrees)
displacement	Displacement (cu. in.)
gear_ratio	Gear Ratio
foreign	Foreign

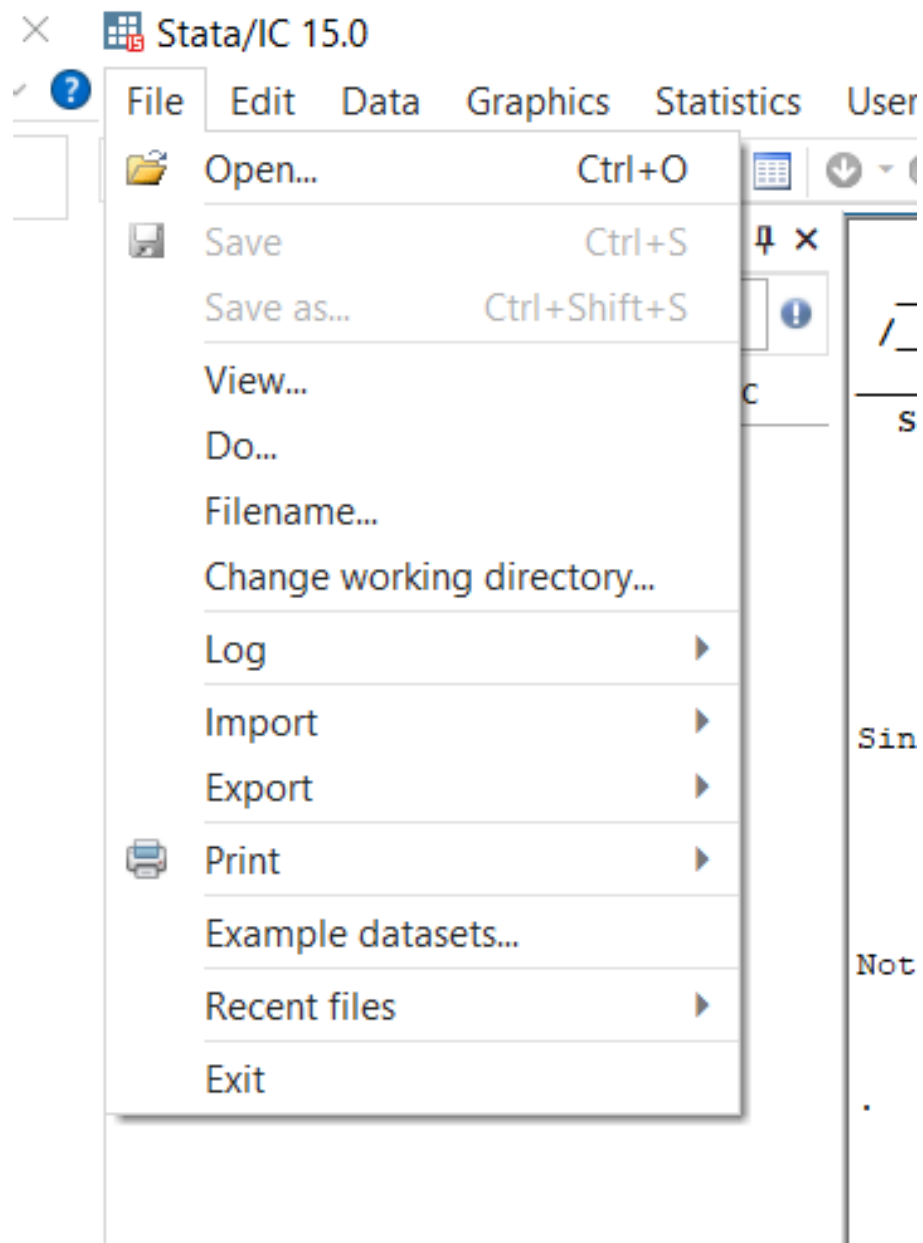
**List of all
variables in
data**

Properties

Variable	Type	Total Obs.	Data Name
make	string	74	make
price	float	74	price
mpg	float	74	mpg
rep78	float	74	rep78
headroom	float	74	headroom
trunk	float	74	trunk
weight	float	74	weight
length	float	74	length
turn	float	74	turn
displacement	float	74	displacement
gear_ratio	float	74	gear_ratio
foreign	float	74	foreign

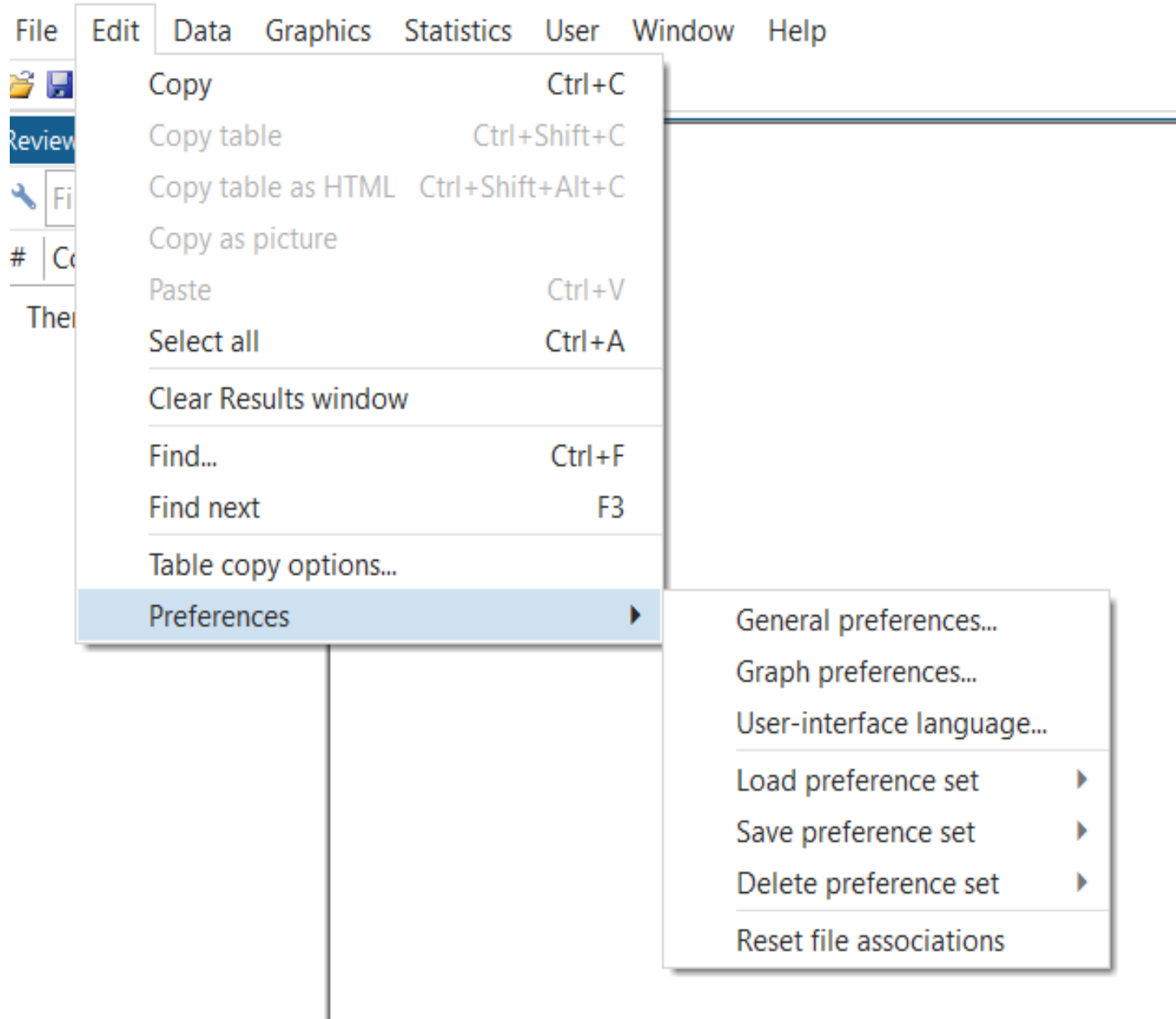
Variables: 3.11K
 Observations: 64M
 Sorted by: foreign

**Type of
variable, total
obs., data name
etc.**



Click on **File**

Load data into the working directory from here



Click on **Edit** → **Preferences**
Here you can change the appearance of your interface, its color etc

Stata/IC 15.0 - C:\Users\shweta.gupta\Desktop\stata1

File Edit **Data** Graphics Statistics User Win



Review

Filter con

Commar

There are r
sho

- Describe data ▶
- Data Editor ▶
- Create or change data ▶
- Variables Manager
- Data utilities ▶
- Sort
- Combine datasets ▶
- Matrices, Mata language
- Matrices, ado language ▶
- ICD codes ▶
- Other utilities ▶

Click on **Data**

View your data, make changes
in variables, work with variables



Review

Filter commands

Command

There are no items
show.

- Twoway graph (scatter, line, etc.)
- Bar chart
- Dot chart
- Pie chart
- Histogram
- Box plot
- Contour plot
- Scatterplot matrix
- Distributional graphs ▶
- Smoothing and densities ▶
- Regression diagnostic plots ▶
- Time-series graphs ▶
- Panel-data line plots
- Survival analysis graphs ▶
- ROC analysis ▶
- Multivariate analysis graphs ▶
- Quality control ▶
- More statistical graphs ▶
- Table of graphs
- Manage graphs ▶
- Change scheme/size

Click on **Graphics**
Create various kinds of graphs
and charts

There are no items to show.

- Summaries, tables, and tests
- Linear models and related
- Binary outcomes
- Ordinal outcomes
- Categorical outcomes
- Count outcomes
- Fractional outcomes
- Generalized linear models
- Time series
- Multivariate time series
- Spatial autoregressive models
- Longitudinal/panel data
- Multilevel mixed-effects models
- Survival analysis
- Epidemiology and related
- Endogenous covariates
- Sample-selection models
- Treatment effects
- SEM (structural equation modeling)
- LCA (latent class analysis)
- FMM (finite mixture models)
- IRT (item response theory)
- Survey data analysis
- Multiple imputation
- Nonparametric analysis
- Multivariate analysis
- Exact statistics
- Resampling
- Power and sample size
- Bayesian analysis
- Postestimation
- Other

Click on **Statistics**

Do all kinds of analysis like find mean, SD, linear regression, etc.



Review [minimize] [maximize] [close]

Filter commands [info]

#	Command	_rc
---	---------	-----

There are no items to show.

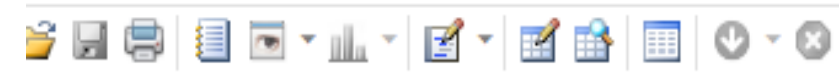
- Command Ctrl+1
- Results Ctrl+2
- Review Ctrl+3
- Variables Ctrl+4
- Properties Ctrl+5
- Graph ▶
- Viewer ▶
- Data Editor Ctrl+8
- Do-file Editor ▶
- Variables Manager

Click on **Window**

Open, view, hide various windows in STATA

Stata/IC 15.0

File Edit Data Graphics Statistics User Window Help



Review

Filter commands

#	Command	_rc
---	---------	-----

There are no items to show.

Help

PDF documentation

Advice

Contents

Search...

Stata command...

News

Resources

SJ and community-contributed commands

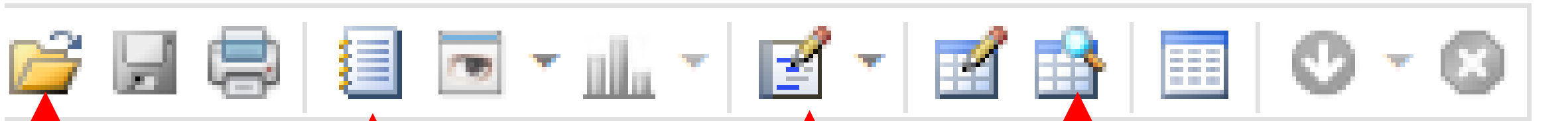
What's new?

Check for updates

About Stata

Click on **Help**

Take help on command, syntax, etc.



Open a file

Save the data in STATA format- dta

Print results

Create a new log file

View the graph created by you




Create a new Do file

Data editor-view and edit data

Data browser-only view data

View variable list and their properties

Types of files generated by STATA

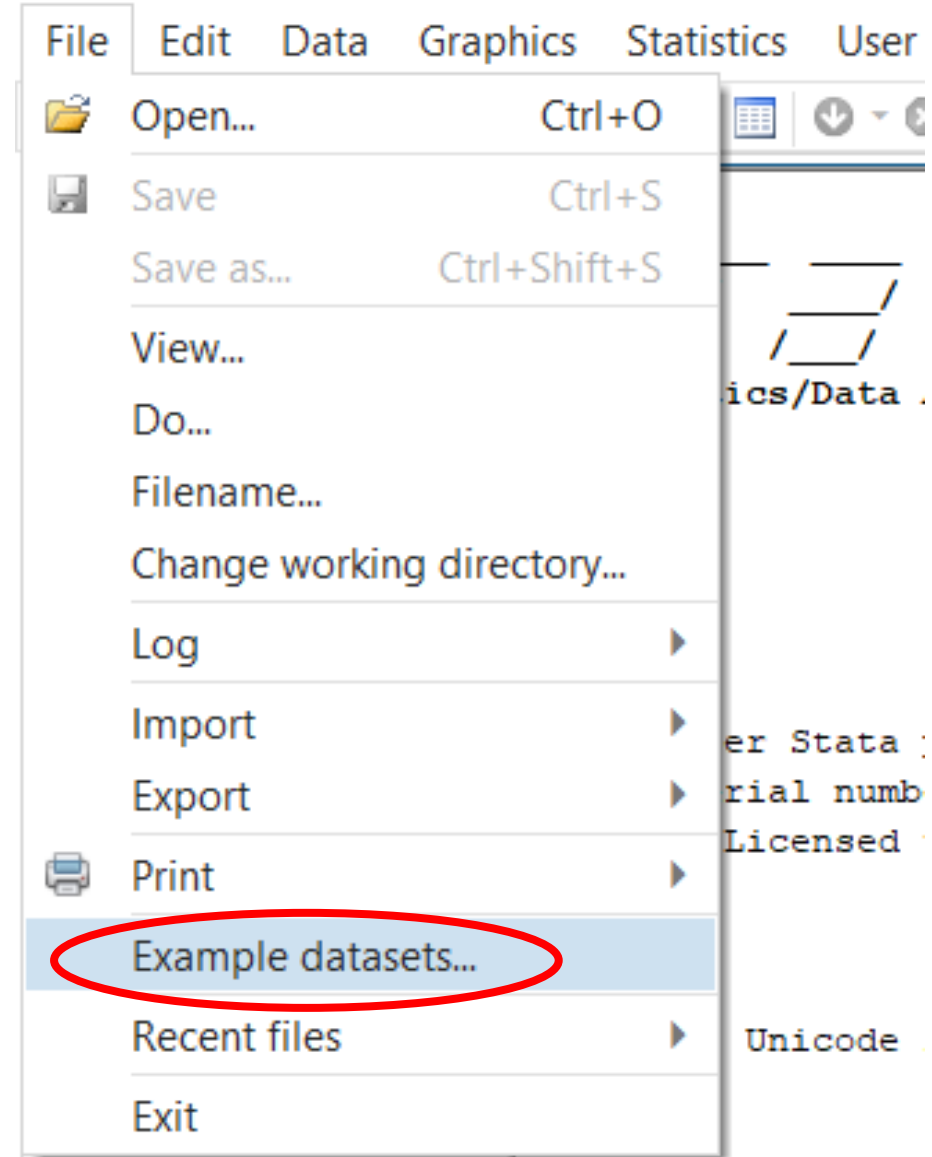
Name	Date modified	Type	Size
 data.dta	5/24/2017 10:10 A...	Stata Dataset	7 KB
 do file.do	3/30/2022 3:31 PM	Stata Do-file	1 KB
 log file.smcl	3/30/2022 3:34 PM	Stata SMCL document	1 KB

Step 1: Load data into STATA

1. Using Example data sets already stored in STATA

- Click on **File** → **Example Datasets**

Stata/IC 15.0



- Click on **Example Datasets** installed with STATA

The screenshot shows the STATA software interface. At the top, there is a menu bar with 'Graphics', 'Statistics', 'User', 'Window', and 'Help'. Below the menu bar is a toolbar with various icons. The main window is titled 'Viewer - help dta_contents' and has a menu bar with 'File', 'Edit', 'History', and 'Help'. The address bar shows 'help dta_contents'. The main content area displays the help page for 'help dta_contents'. The page has a 'Title' section with the text '[U] 1.2.2 Example datasets' and a 'Description' section with the text 'Example datasets installed with Stata'. The text 'Example datasets installed with Stata' is highlighted with a red box. Below the description, there is a paragraph: 'This page contains links enabling you to describe or use the datasets that were installed with Stata.' and a link 'Stata 15 manual datasets'. At the bottom right of the window, there is a status bar with the text 'CAP NUM OVR'.

- Click on **use** to load data

ata Graphics Statistics User Window Help

Viewer - help dta_examples

File Edit History Help

help dta_examples

help dta_examples x

Example datasets installed with Stata

The datasets listed here are installed with Stata. You can also see the [datasets used in the Stata documentation](#) that are available in the manual. The manual title is listed as a link that will take you to the list of datasets in the manual.

Notes:

auto.dta	use	describe
auto2.dta	use	describe
autornd.dta	use	describe
bplong.dta	use	describe
bpwide.dta	use	describe
cancer.dta	use	describe
census.dta	use	describe
citytemp.dta	use	describe
citytemp4.dta	use	describe
educ99gdp.dta	use	describe
gnp96.dta	use	describe
lifeexp.dta	use	describe
network1.dta	use	describe
network1a.dta	use	describe
nlsw88.dta	use	describe
nlswid1.dta	use	describe
pop2000.dta	use	describe
sandstone.dta	use	describe
sp500.dta	use	describe
surface.dta	use	describe
tsline1.dta	use	describe
tsline2.dta	use	describe
uslifeexp.dta	use	describe
uslifeexp2.dta	use	describe
voter.dta	use	describe
xtline1.dta	use	describe

#	Command	_rc
1	sysuse auto.dta	

```
(R)
-----
  /  /  /  /  /
 /  /  /  /  /
-----
Statistics/Data Analysis      15.0

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)

Single-user Stata perpetual license:
  Serial number: 301506215585
  Licensed to:  www.Downloadly.ir
                Iran Will Defeat US

Notes:
  1. Unicode is supported; see help unicode\_advice.

. sysuse auto.dta
(1978 Automobile Data)

.
```

Name	Label
make	Make
price	Price
mpg	Mileage
rep78	Repair
headroom	Headroom
trunk	Trunk space
weight	Weight
length	Length
turn	Turn circle
displacement	Displacement
gear_ratio	Gear Ratio
foreign	Car type

Variables
Name
Label
Type
Format
Value label
Notes

Data
Filename
Label
Notes



Data is loaded into STATA

Erase the existing data

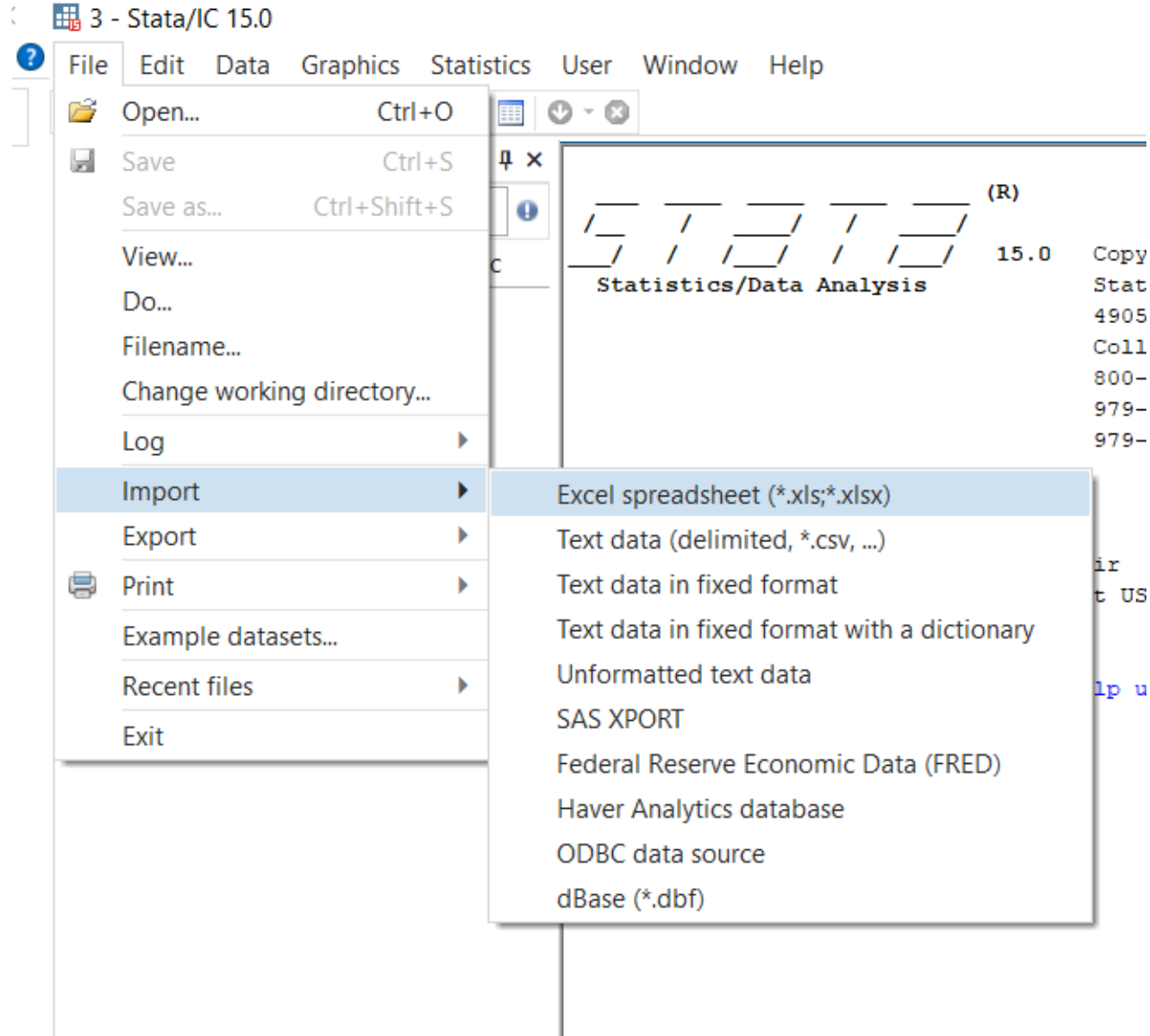
```
Command  
clear all|
```

In the command window
type:

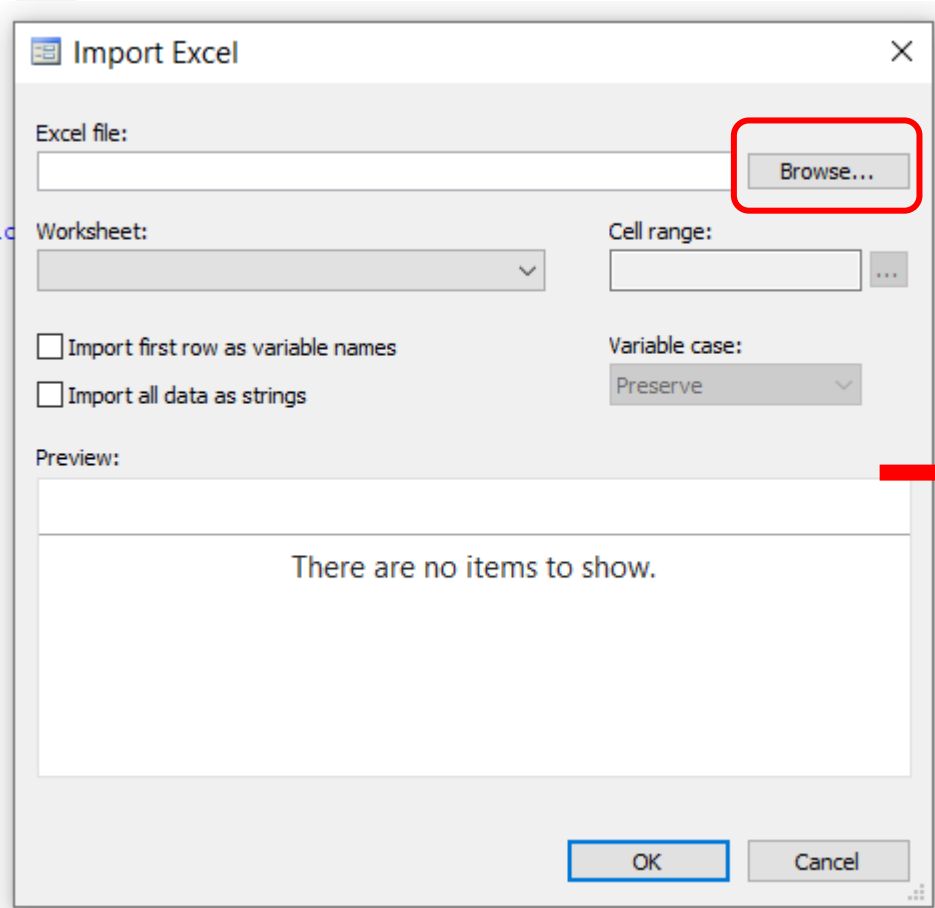
clear all

Press enter

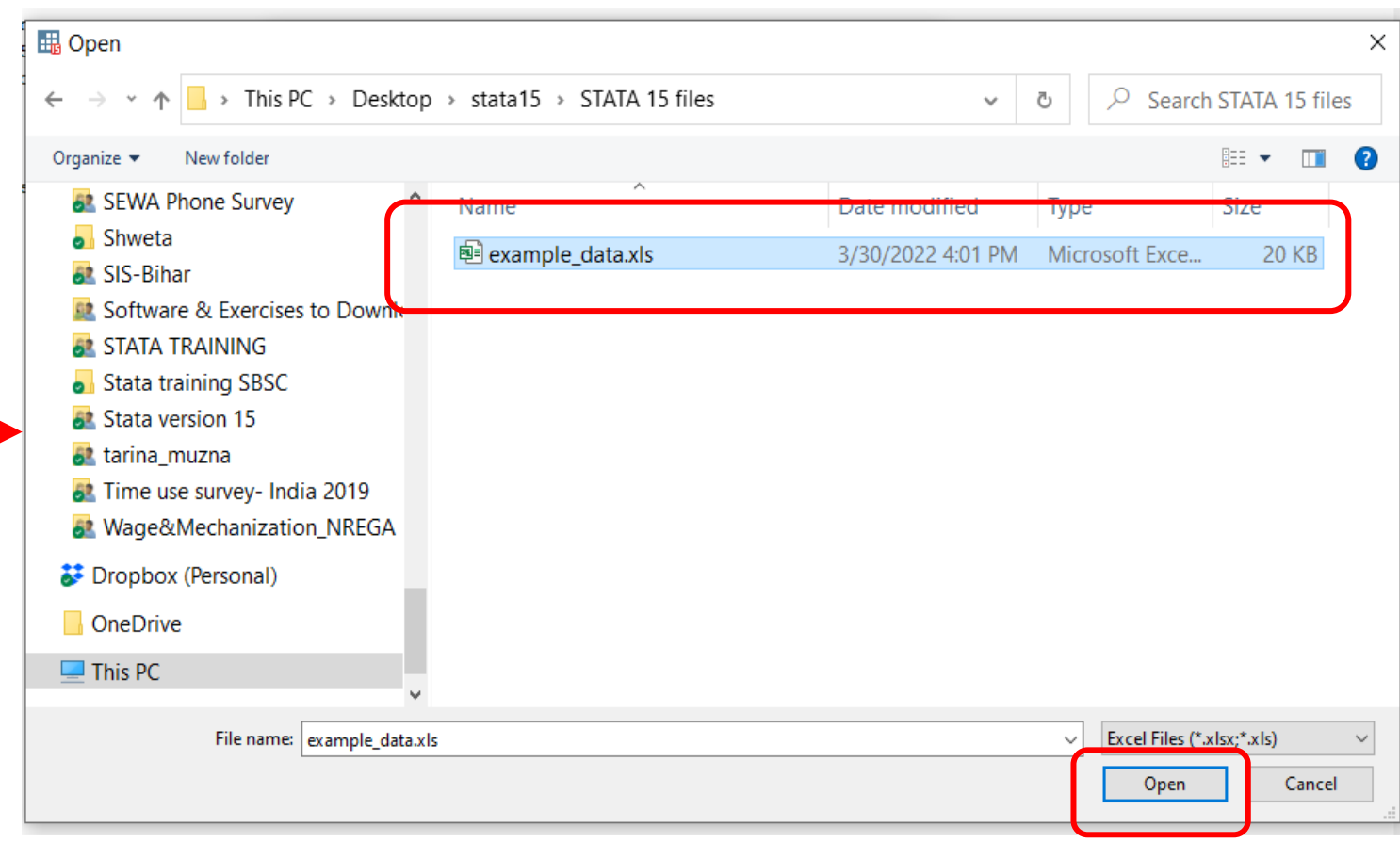
2. Import data in non-STATA format; eg- excel file



Click on **File** → **Import** → **Excel spreadsheet**



Click on **Browse**



Select the excel file
and click **Open**

Select **Import first row as variable names**

ic

Import Excel

Excel file:
C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\example_data.xls

Worksheet:
Sheet1 A1:L75

Cell range:
A1:L75

Import first row as variable names
 Import all data as strings

Variable case:
Preserve

Preview: (showing rows 2-51 of 75)

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement
2	AMC Concord	4099	22	3	2.5	11	2930	186	40	121
3	AMC Pacer	4749	17	3	3	11	3350	173	40	258
4	AMC Spirit	3799	22	.	3	12	2640	168	35	121
5	Buick Century	4816	20	3	4.5	16	3250	196	40	196
6	Buick Electra	7827	15	4	4	20	4080	222	43	350
7	Buick LeSabre	5788	18	3	4	21	3670	218	43	231
8	Buick Opel	4453	26	.	3	10	2230	170	34	304
9	Buick Regal	5189	20	3	2	16	3280	200	42	196
10	Buick Riviera	10372	16	3	3.5	17	3880	207	43	231
11	Buick Skylark	4082	19	3	3.5	13	3400	200	42	231

OK Cancel

Click Ok



Review

Filter command
Command
1 import ex...

```
(R)
-----
Statistics/Data Analysis 15.0 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)

Single-user Stata perpetual license:
  Serial number: 301506215585
  Licensed to: www.Downloadly.ir
             Iran Will Defeat US

Notes:
  1. Unicode is supported; see help unicode_advice.

. import excel "C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\example_data.xls", sheet("Sheet1") firstrow
```

Data is loaded

Command

Variables

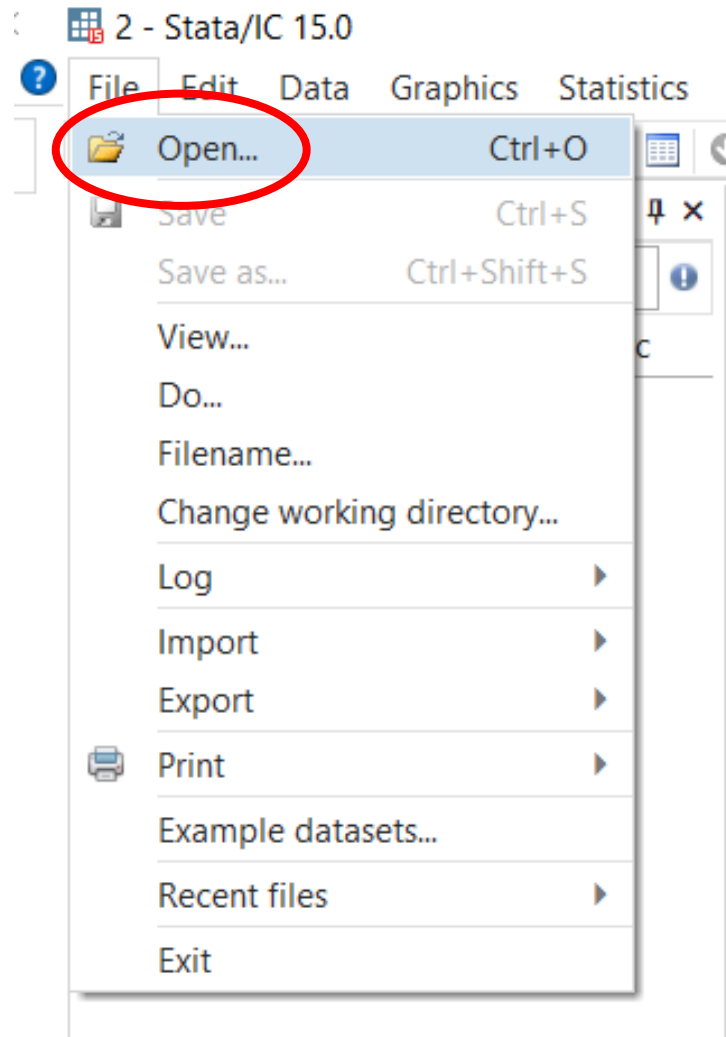
Filter variables here

Name	Label
make	make
price	price
mpg	mpg
rep78	rep78
headroom	headroom
trunk	trunk
weight	weight
length	length
turn	turn
displacement	displacement
gear_ratio	gear_ratio

Properties

Variables	
Name	Label
Type	
Format	
Value label	
Notes	
Data	
Filename	
Label	
Notes	
Variables	12
Observations	74
Size	2.02K

3. Use a STATA data file- dta format

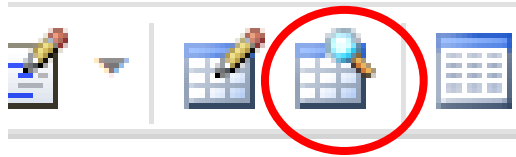


Click on **Open**

Select folder and choose dta file

Exercise: Open auto.dta file from Example datasets
Use this file for now

Step 2: Viewing data



- Click on Data browser



Data Editor (Browse) - [auto.dta]

File Edit View Data Tools

make[1] AMC Concord

	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	2.56	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

Variables

Filter variables here

<input checked="" type="checkbox"/>	Name	Label
<input checked="" type="checkbox"/>	make	Make and Mo...
<input checked="" type="checkbox"/>	price	Price
<input checked="" type="checkbox"/>	mpg	Mileage (mpg)
<input checked="" type="checkbox"/>	rep78	Repair Record...
<input checked="" type="checkbox"/>	headroom	Headroom (in.)
<input checked="" type="checkbox"/>	trunk	Trunk space (c...
<input checked="" type="checkbox"/>	weight	Weight (lbs.)
<input checked="" type="checkbox"/>	length	Length (in.)

Variables Snapshots

Properties

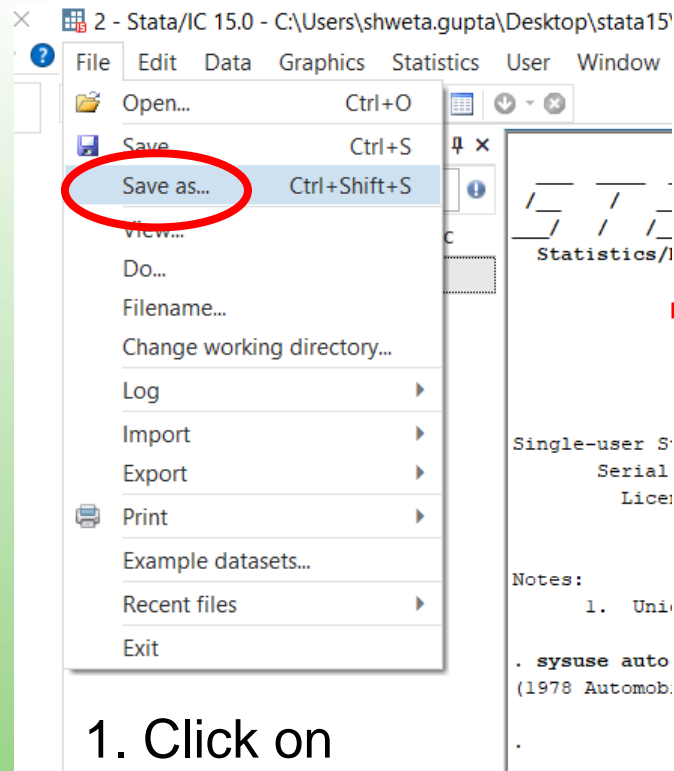
Variables	
Name	make
Label	Make and Mo
Type	str18
Format	%-18s
Value label	
Notes	

Data

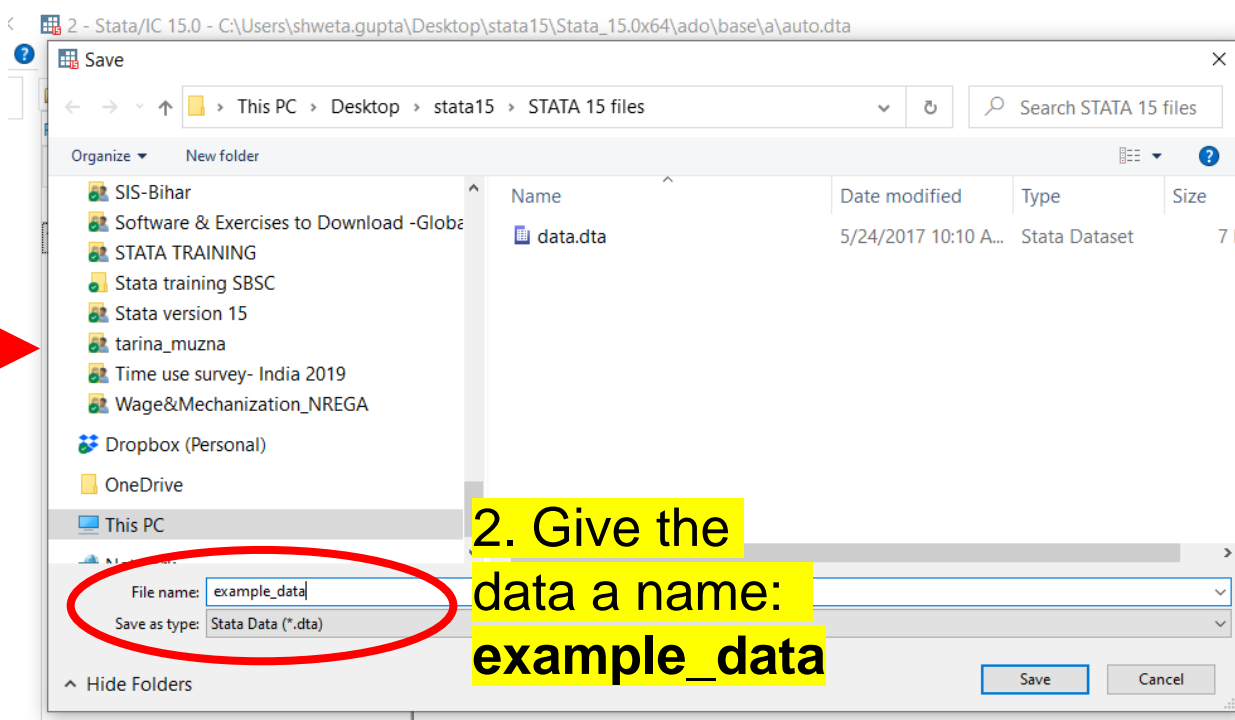
Filename	auto.dta
Label	1978 Automol

Ready Length: 18 Vars: 12 Order: Dataset Obs: 74 Filter: Off Mode: Browse CAP NUM

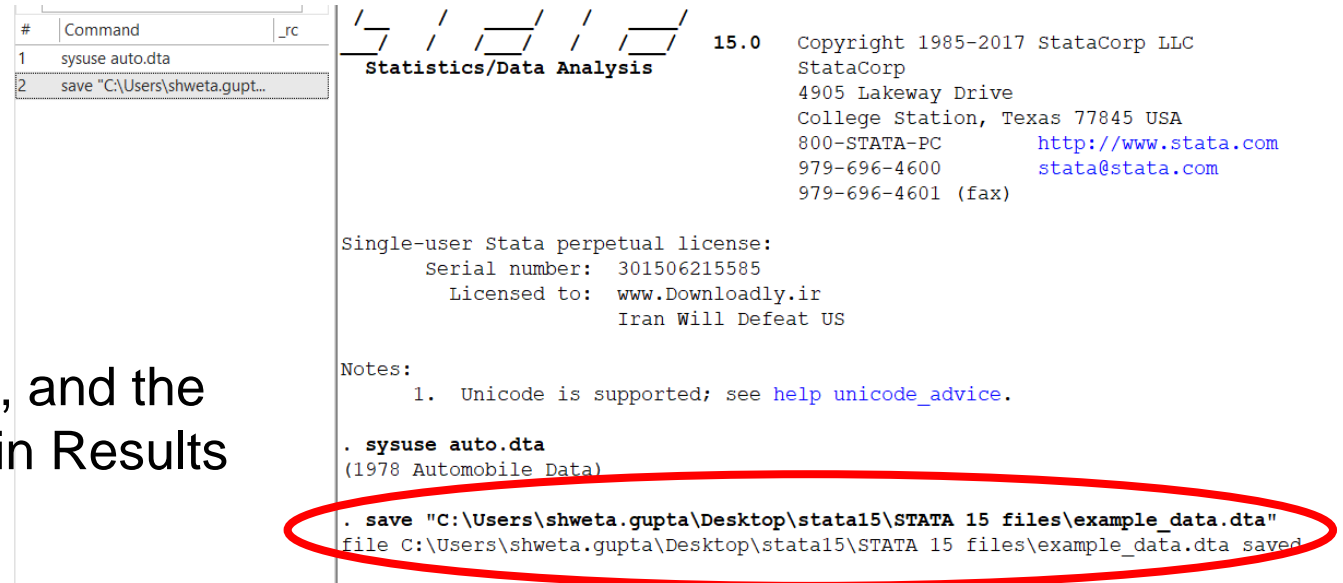
Step 3: Saving this data as dta file



1. Click on **File** → **Save as**

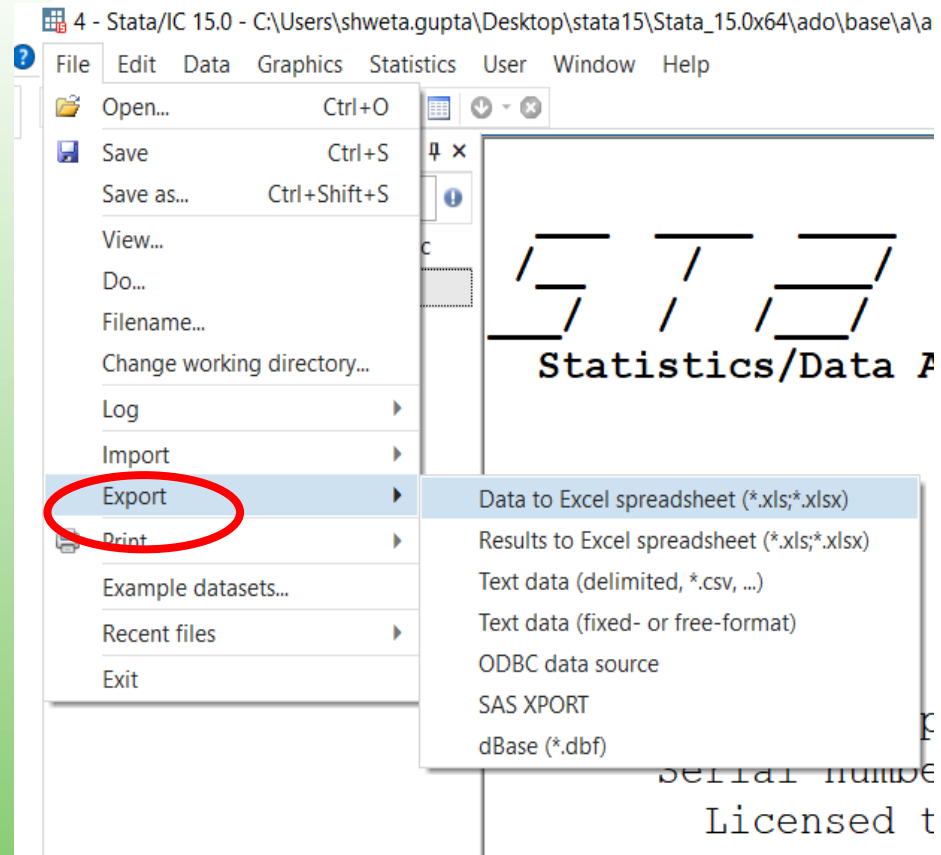


2. Give the data a name: **example_data**

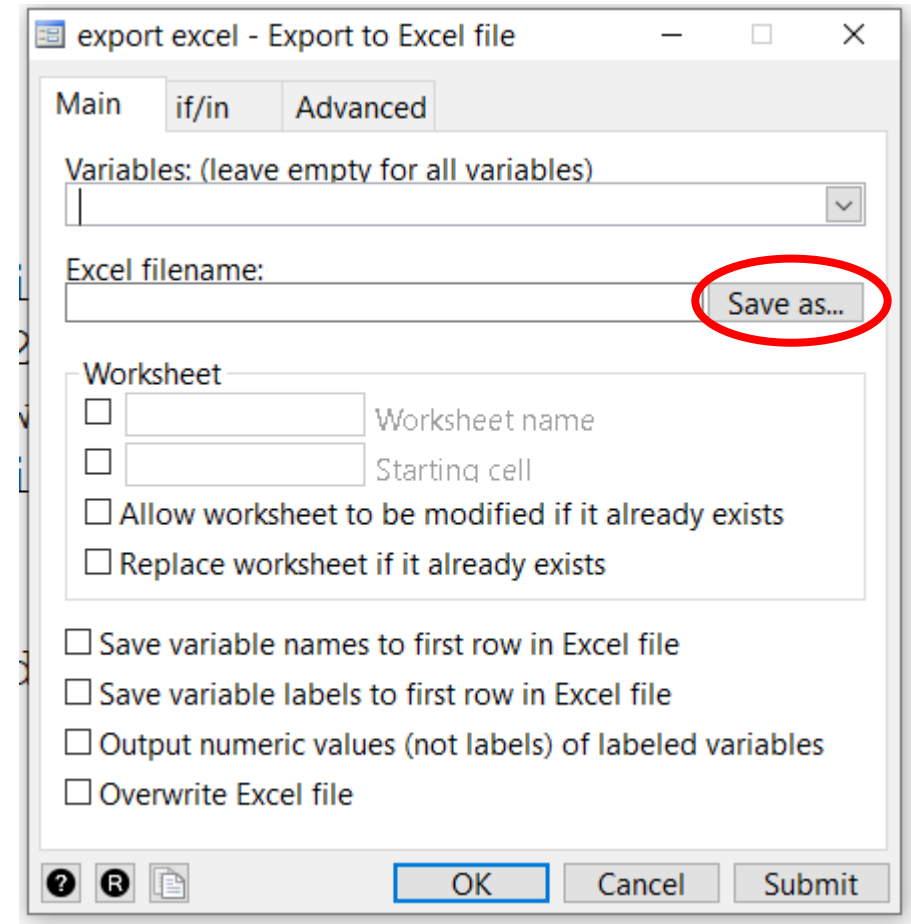


3. The file is saved, and the result is displayed in Results window

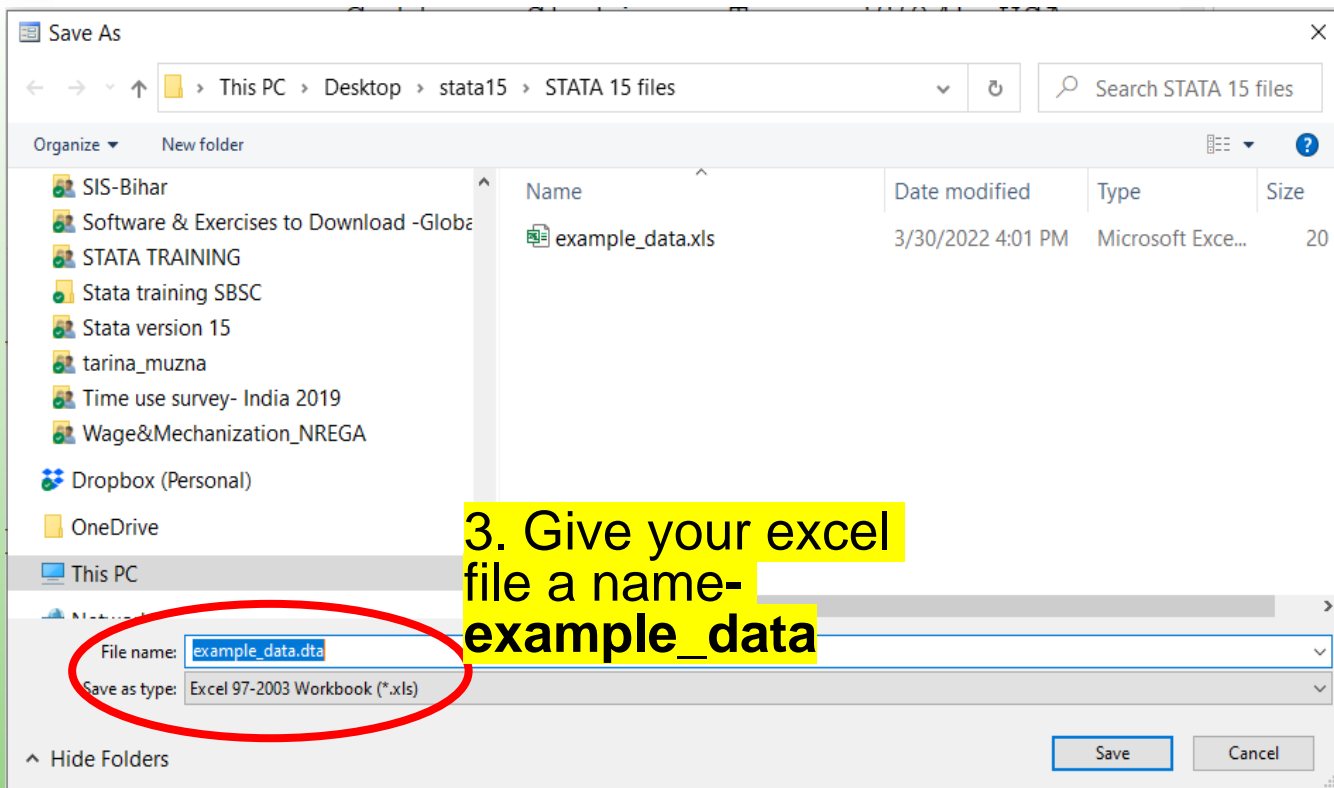
Step 3: Saving this data as excel file



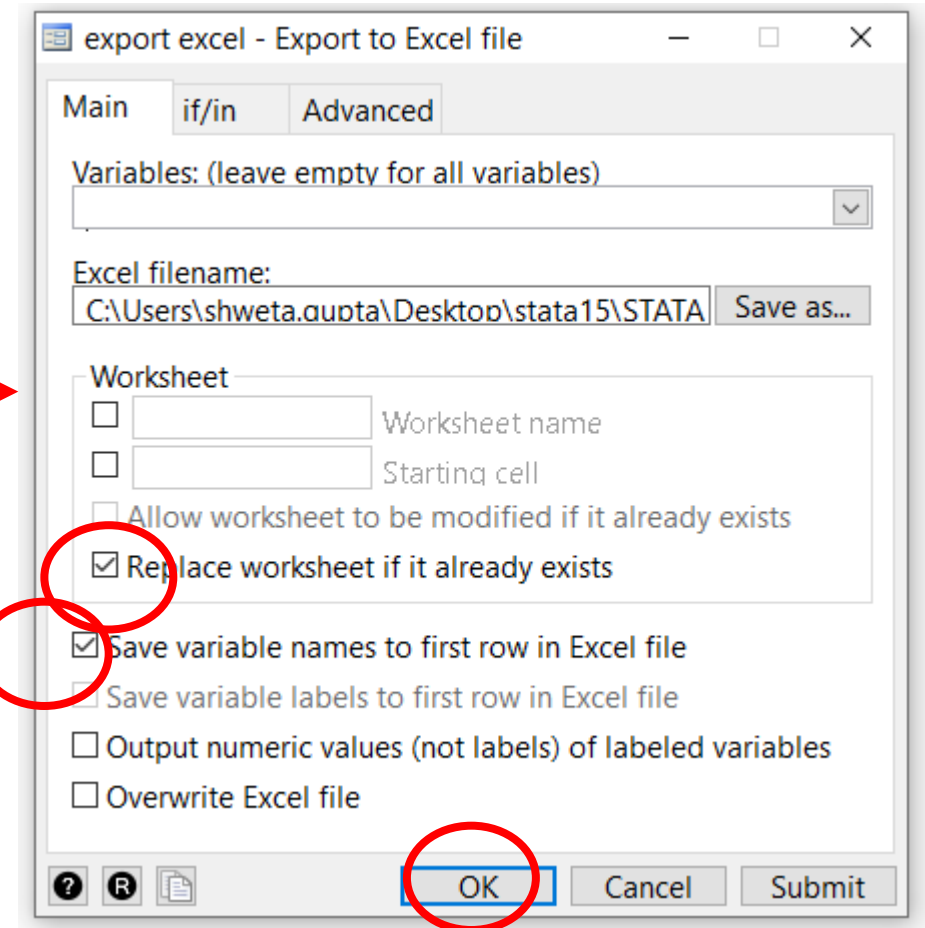
1. Click on **Export** →
Data to Excel
spreadsheet



2. Click on **Save as..**



3. Give your excel file a name-
example_data



4. Check the box of above 2 options. Then click OK

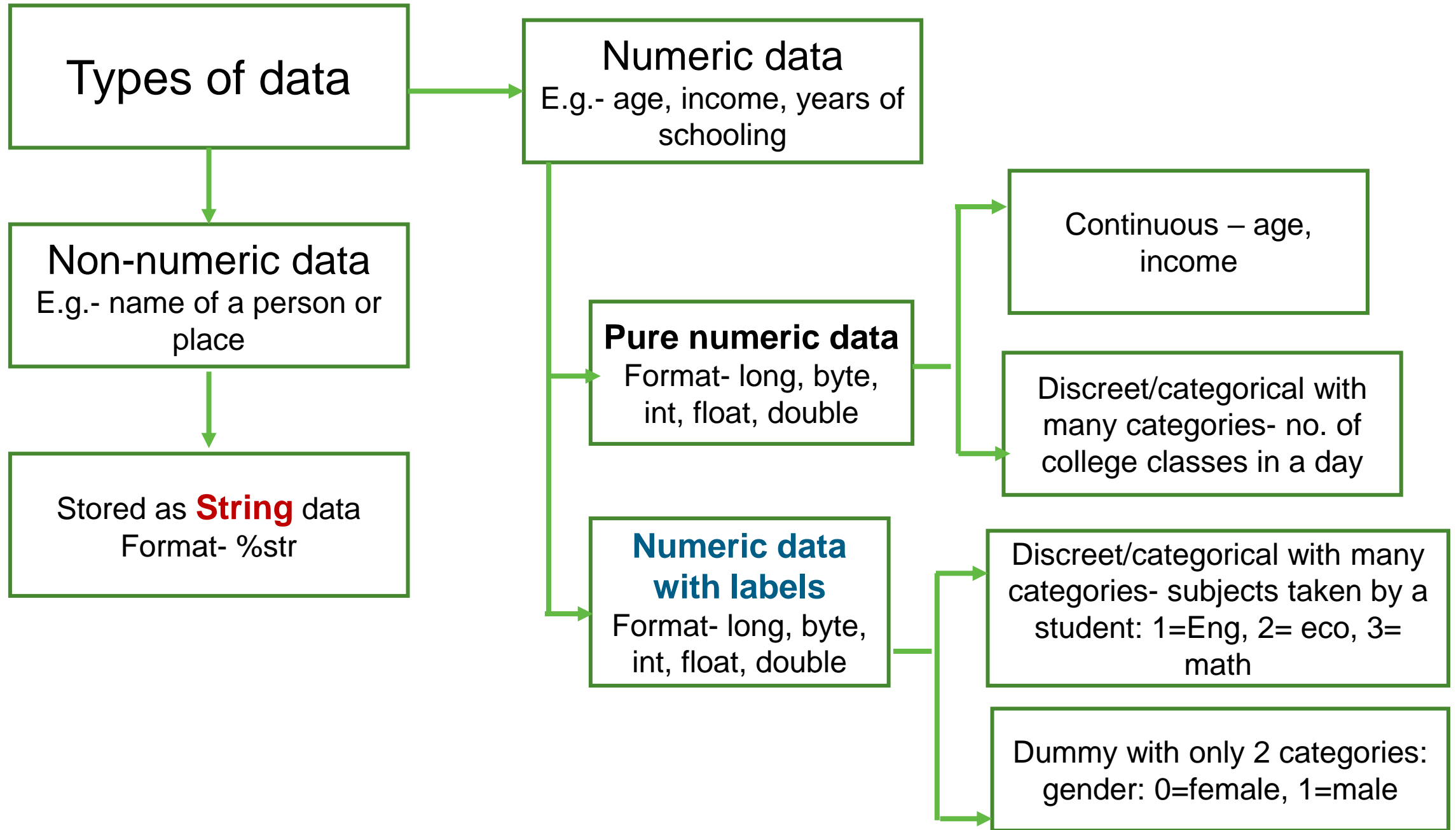


5. Data is saved in xls format

```
. sysuse auto.dta
(1978 Automobile Data)

export excel using "C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\example_data.xls", sheetreplace firstrow(variables)
file C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\example_data.xls saved
```

Part 2: Understanding your data



sysuse auto.dta

Name of variables

Selected cell value is shown at top

Observation no.

	mpg[4]	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

String data (non-numeric)

Pure numeric data

Numeric data with label attached



View data

1. Browse

browse → view entire data

br → in short, view entire data

br make price mpg → view only mentioned variables



Click on data browser

2. Count no. of observations

count

Describe data

3. Obtain the list of all variable names, their type, their label, data name

describe

des

d

d make price mpg

```
. des
Contains data from C:\Users\shweta.gupta\Desktop\stata15\Stata_15.0x64\ado\base/a/auto.dta
  obs:          74                1978 Automobile Data
  vars:         12                13 Apr 2016 17:45
  size:        3,182              (_dta has notes)
-----
      storage  display  value
variable name  type    format  label    variable label
-----
make           str18   %-18s   Make and Model
price          int     %8.0gc  Price
mpg            int     %8.0g   Mileage (mpg)
rep78          int     %8.0g   Repair Record 1978
headroom       float   %6.1f   Headroom (in.)
trunk          int     %8.0g   Trunk space (cu. ft.)
weight         int     %8.0gc  Weight (lbs.)
length         int     %8.0g   Length (in.)
turn           int     %8.0g   Turn Circle (ft.)
displacement   int     %8.0g   Displacement (cu. in.)
gear_ratio     float   %6.2f   Gear Ratio
foreign        byte    %8.0g   origin   Car type
-----
Sorted by: foreign
```

Storage type and display format

Storage type:

- `str18` → string variable with 18 characters allowed
- `int`, `float`, `byte` → numeric variables

Display format:

- `%-18s` → string variable, 18 characteristics, left-justified
- `%6.1f` → fixed format numeric variable, 6 characters, 1 decimal, right justified
- `%8.0g` → general format numeric variable, 8 characters, no decimal, right justified

Summarize data

4. Summarize data to get mean, SD, etc

sum

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

sum make price mpg

```
. sum make price mpg
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41

Summarize in detail

5. Detailed summary of a variable sum mpg, d

```
. sum mpg, d
```

			Mileage (mpg)	
Percentiles			Smallest	
1%	12	12		
5%	14	12		
10%	14	14	Obs	74
25%	18	14	Sum of Wgt.	74
50%	20		Mean	21.2973
		Largest	Std. Dev.	5.785503
75%	25	34	Variance	33.47205
90%	29	35	Skewness	.9487176
95%	34	35	Kurtosis	3.975005
99%	41	41		

Compare it with sum mpg

```
. sum mpg
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mpg	74	21.2973	5.785503	12	41

See data in Results window

6. list command

list →view entire data

list price →view all observations of price

list in 1/10 →view only 1st 10 observations for all variables

list price mpg in 1/10 →view only 1st obs. of price and mpg

```
. list in 1/10
```

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displ~t	gear_r~o	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

Sort variables

6. sort command

`sort make` → sort alphabetically over make

`sort price` → sort from smallest to largest over price

`sort foreign make` → sort by foreign first, then in each category of foreign, sort alphabetically by make

Go to data browser to see how data gets sorted from above

Change order of variables in dataset

7. order command

`order make foreign` → after make, put foreign

`order foreign, first` → bring foreign in the beginning

`order price, last` → push price to the last

`order price, a(foreign)` → put price after foreign

`order price, b(foreign)` → bring price before foreign

Go to data browser to see how variables get ordered

Error in STATA

```
. sort  
varlist required  
r(100);
```

```
. order  
varlist required  
r(100);
```

Instructions to
remove error

```
. sort data  
variable data not found  
r(111);
```

```
. list make in 1 to 10  
invalid 'to'  
r(198);
```

```
. list mpg for 1/10  
1 invalid name  
r(198);
```

Best way to get rid
of error:

```
help sort  
help command
```

Tabulate variables

7. tabulate command

a) One-way tabulate

tab rep78

```
. tab rep78
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.90	2.90
2	8	11.59	14.49
3	30	43.48	57.97
4	18	26.09	84.06
5	11	15.94	100.00
Total	69	100.00	

Frequency distribution of a var

tab rep78, sort

```
. tab rep78, sort
```

Repair Record 1978	Freq.	Percent	Cum.
3	30	43.48	43.48
4	18	26.09	69.57
5	11	15.94	85.51
2	8	11.59	97.10
1	2	2.90	100.00
Total	69	100.00	

Frequency sorted from high to low

Tabulate variables

7. tabulate command

a) One-way tabulate

`tab foreign`

```
. tab foreign
```

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

`tab foreign, nolabel`

```
. tab foreign, nolabel
```

Car type	Freq.	Percent	Cum.
0	52	70.27	70.27
1	22	29.73	100.00
Total	74	100.00	

Tabulate variables

7. tabulate command

b) Two-way tabulate

`tab rep78 foreign`

```
. tab rep78 foreign
```

Repair Record 1978	Car type		Total
	Domestic	Foreign	
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

`tab rep78 foreign, nolabel`

```
. tab rep78 foreign, nol
```

Repair Record 1978	Car type		Total
	0	1	
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

Tabulate variables

7. tabulate command

b) Two-way tabulate- options

`tab rep78 foreign, col`

Repair Record 1978	Car type		Total
	Domestic	Foreign	
1	2 4.17	0 0.00	2 2.90
2	8 16.67	0 0.00	8 11.59
3	27 56.25	3 14.29	30 43.48
4	9 18.75	9 42.86	18 26.09
5	2 4.17	9 42.86	11 15.94
Total	48 100.00	21 100.00	69 100.00

`tab rep78 foreign, row`

Repair Record 1978	Car type		Total
	Domestic	Foreign	
1	2 100.00	0 0.00	2 100.00
2	8 100.00	0 0.00	8 100.00
3	27 90.00	3 10.00	30 100.00
4	9 50.00	9 50.00	18 100.00
5	2 18.18	9 81.82	11 100.00
Total	48 69.57	21 30.43	69 100.00

Tabulate variables

7. tabulate command

b) Two-way tabulate- options

```
tab rep78 foreign, col nofreq
```

```
. tab rep78 foreign, col nofreq
```

Repair Record 1978	Car type		Total
	Domestic	Foreign	
1	4.17	0.00	2.90
2	16.67	0.00	11.59
3	56.25	14.29	43.48
4	18.75	42.86	26.09
5	4.17	42.86	15.94
Total	100.00	100.00	100.00

Tabulate variables

8. table command

a) one-way table

`table rep78`

```
. table rep78
```

Repair Record 1978	Freq.
1	2
2	8
3	30
4	18
5	11

`table rep78, row`

```
. table rep78, row
```

Repair Record 1978	Freq.
1	2
2	8
3	30
4	18
5	11
Total	69

Tabulate variables

8. table command

b) two-way table

`table rep78 foreign`

```
. table rep78 foreign
```

Repair Record 1978	Car type	
	Domestic	Foreign
1	2	
2	8	
3	27	3
4	9	9
5	2	9

`table rep78 foreign, row`

```
. table rep78 foreign, row
```

Repair Record 1978	Car type	
	Domestic	Foreign
1	2	
2	8	
3	27	3
4	9	9
5	2	9
Total	48	21

`table rep78 foreign, row col`

```
. table rep78 foreign, row col
```

Repair Record 1978	Car type		Total
	Domestic	Foreign	
1	2		2
2	8		8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

Tabulate variables

8. table command

c) Three-way table

`table headroom rep78 foreign`

```
. table headroom rep78 foreign, col row
```

Headroom (in.)	Car type and Repair Record 1978											
	Domestic						Foreign					
	1	2	3	4	5	Total	1	2	3	4	5	Total
1.5	1			1		2				1		1
2.0	1	3	5		1	10				1	2	3
2.5			3		1	4			2	5	3	10
3.0			3	2		5			1	1	4	6
3.5		1	10	1		12				1		1
4.0		2	3	5		10						
4.5		1	3			4						
5.0		1				1						
Total	2	8	27	9	2	48			3	9	9	21

Tabulate variables

8. table command

c) Table to get some statistics other than frequency

`table foreign, c(mean headroom mean price)`

```
. table foreign, c(mean headroom mean price)
```

Car type	mean(headroom)	mean(price)
Domestic	3.2	6,072.4
Foreign	2.6	6,384.7

`table rep78 foreign, c(mean price)`

```
. table rep78 foreign, c(mean price )
```

Repair Record 1978	Car type	
	Domestic	Foreign
1	4,564.5	
2	5,967.6	
3	6,607.1	4,828.7
4	5,881.6	6,261.4
5	4,204.5	6,292.7

Combine tabulate and summarize

9. tab, sum() command

```
tab foreign, sum(price)
```

```
. tab foreign, sum(price)
```

Car type	Summary of Price		Freq.
	Mean	Std. Dev.	
Domestic	6,072.423	3,097.104	52
Foreign	6,384.682	2,621.915	22
Total	6,165.257	2,949.496	74

Show results by type

10. bysort command

bysort foreign: sum(price)

```
. bysort foreign: sum price
```

```
-> foreign = Domestic
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	52	6072.423	3097.104	3291	15906

```
-> foreign = Foreign
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	22	6384.682	2621.915	3748	12990

bysort foreign: tab rep78

```
. bysort foreign: tab rep78
```

```
-> foreign = Domestic
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	4.17	4.17
2	8	16.67	20.83
3	27	56.25	77.08
4	9	18.75	95.83
5	2	4.17	100.00
Total	48	100.00	

```
-> foreign = Foreign
```

Repair Record 1978	Freq.	Percent	Cum.
3	3	14.29	14.29
4	9	42.86	57.14
5	9	42.86	100.00
Total	21	100.00	

Exercise: Try bysort with count and tab,sum

If command

11. Can be combined with all commands to execute a command based on a condition
command if (condition), options

```
br if price==6486
br if price>2000
br if price<=2000
br if make=="Plym. Sapporo"
```

```
sum mpg if foreign==0
tab foreign if price>2000, sort
tab foreign rep78 if price>2000, col
```

Combining conditions:

`sum mpg if (foreign==0) & (price>2000)` → summarizes mpg taking only those observations where car is domestic AND price is more than 2000

`sum mpg if (foreign==0) | (gear_ratio==2.5)` → summarizes mpg taking only those observations where EITHER car is domestic OR its gear ratio is 2.5 OR BOTH.

Log file- saving results

- Log file saves all the output from results window in a file.
- Format- SMCL file
- One can open it later to view all work done before.

log using “C:\path\log file.smcl”, replace
(commands)

log close

Show creating log file using Tool bar

Log file- saving results

```
clear all  
log using "C:\path\log file.smcl", replace  
sysuse auto.dta  
sum mpg  
log close
```

Now open the log file to view all history of work.

Result-

```
. clear all

. log using "C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\log file.smcl", replace
```

```
    name: <unnamed>
    log:  C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\log file.smcl
log type: smcl
opened on: 30 Mar 2022, 23:39:48

.

. sysuse auto.dta
(1978 Automobile Data)

.

. sum mpg
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mpg	74	21.2973	5.785503	12	41

```


.

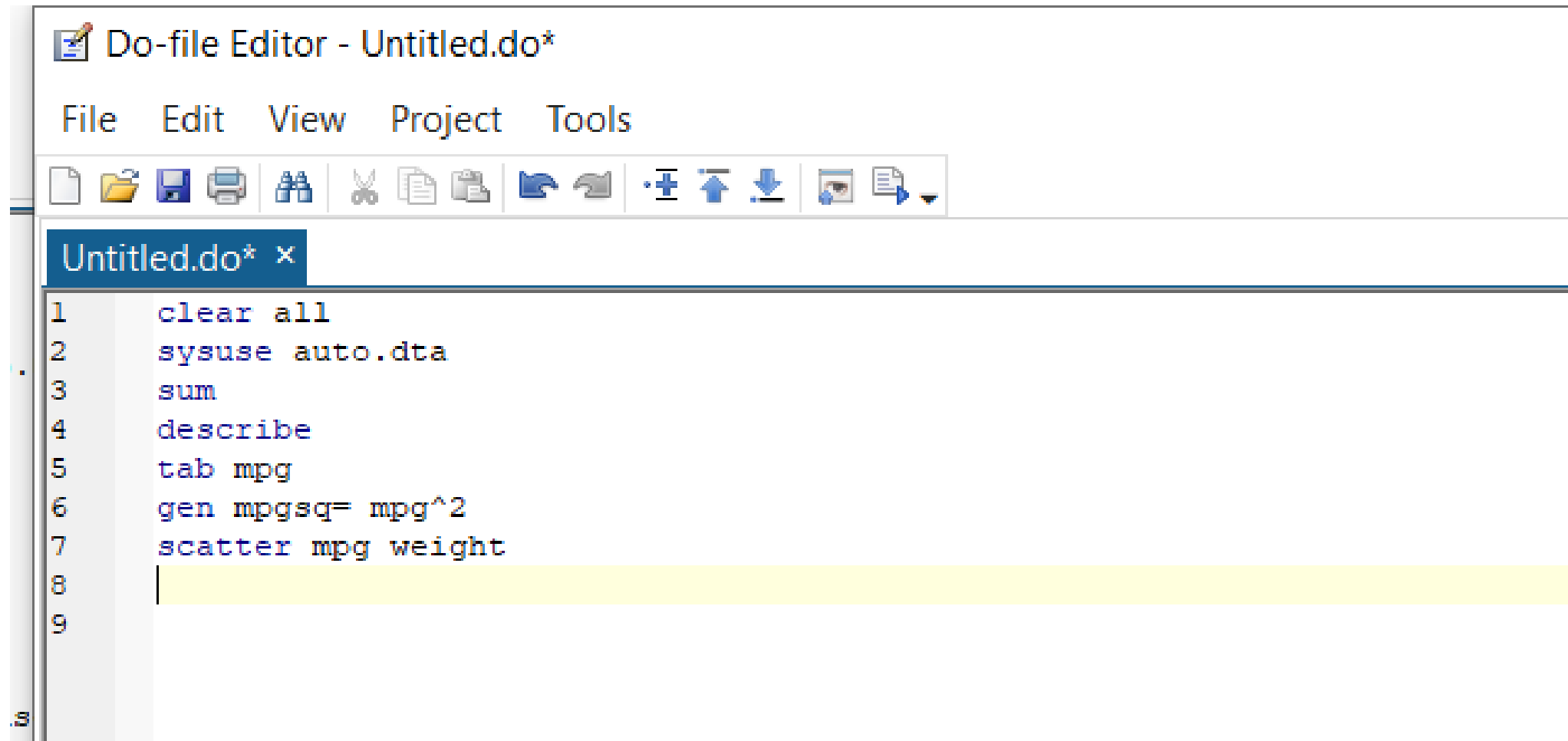
. log close
    name: <unnamed>
    log:  C:\Users\shweta.gupta\Desktop\stata15\STATA 15 files\log file.smcl
log type: smcl
closed on: 30 Mar 2022, 23:39:57
```

Do file- saving commands

- Do file saves all the commands
- Format- .do
- One can use it to rerun all/subset of commands
- Good practice
- Use Toolbar to open a new Do-file



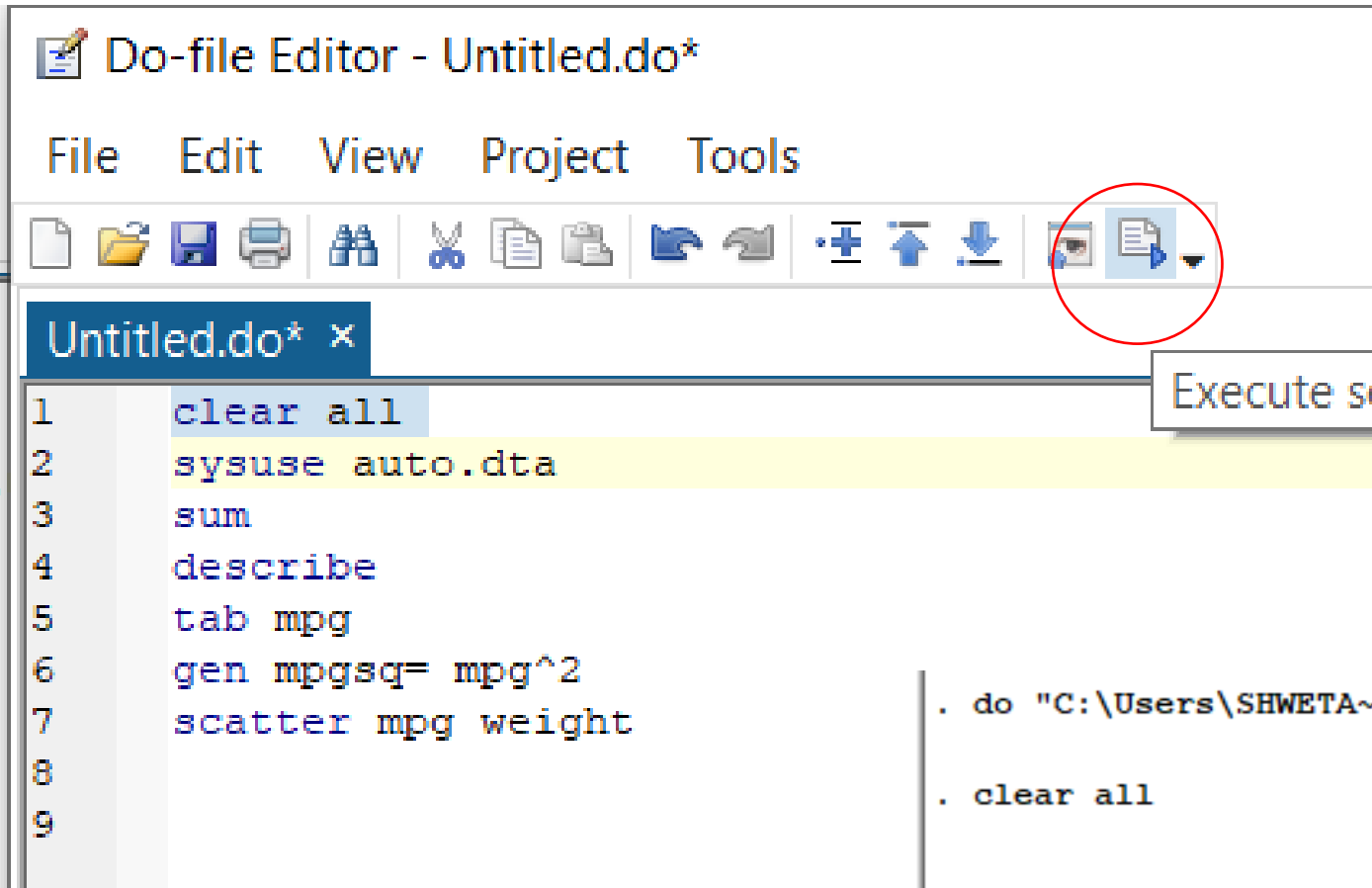
Write commands in Do file



The screenshot shows a software window titled "Do-file Editor - Untitled.do*". The window has a menu bar with "File", "Edit", "View", "Project", and "Tools". Below the menu bar is a toolbar with various icons for file operations and editing. The main area of the window shows a text editor with a blue header bar that says "Untitled.do* x". The text in the editor consists of a list of Stata commands, each on a new line and preceded by a line number from 1 to 9. The commands are: "clear all", "sysuse auto.dta", "sum", "describe", "tab mpg", "gen mpgsq= mpg^2", and "scatter mpg weight". The line number 8 is highlighted in yellow. The line number 9 is also highlighted in yellow. The line number 10 is visible at the bottom left of the editor area.

```
1 clear all
2 sysuse auto.dta
3 sum
4 describe
5 tab mpg
6 gen mpgsq= mpg^2
7 scatter mpg weight
8
9
10
```

Run commands from Do file



Select a line and press the icon in circle

OR

Select a line, then CTRL+ D



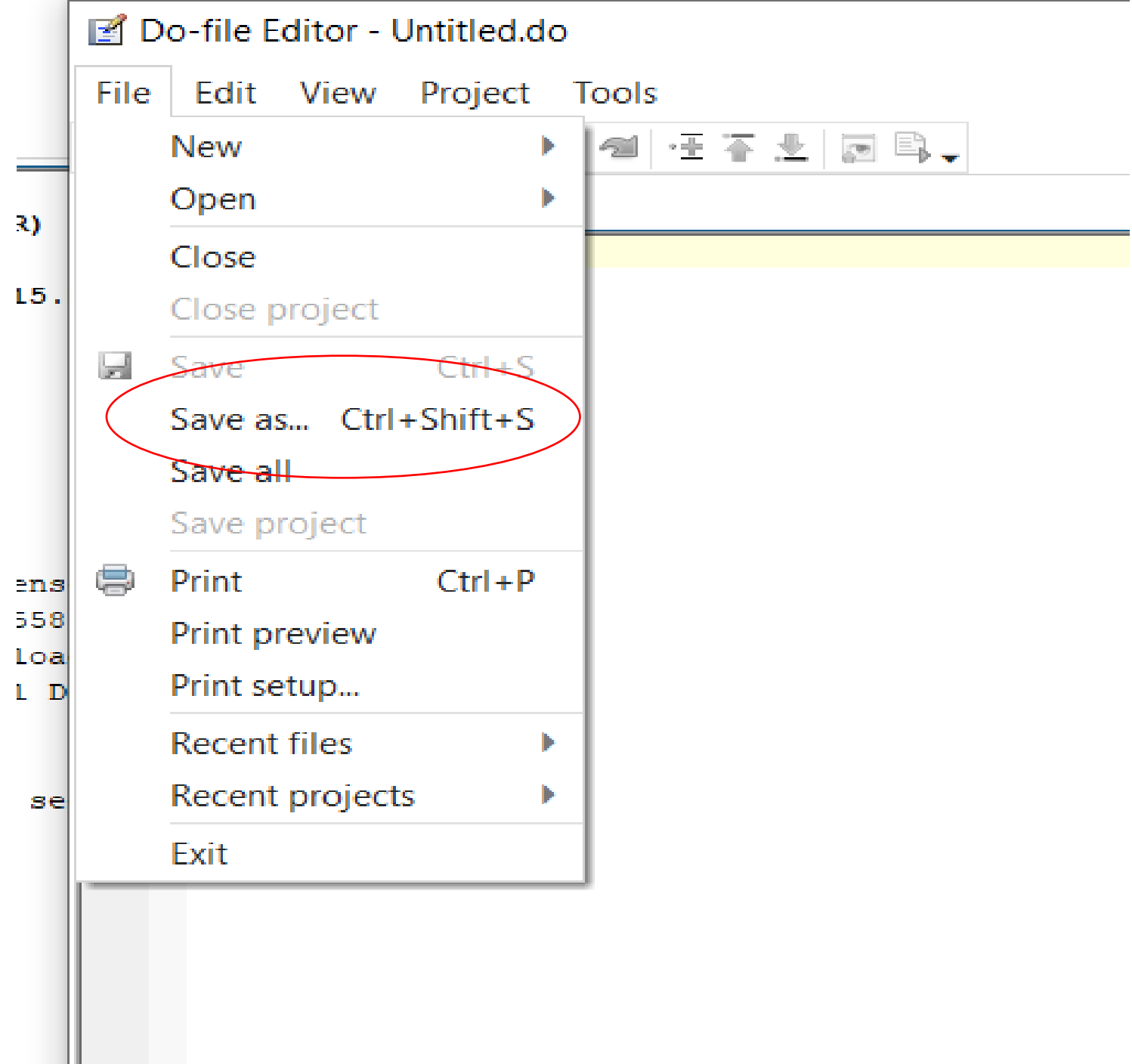
```
. do "C:\Users\SHWETA~1.GUP\AppData\Local\Temp\STD6f48_000000.tmp"

. clear all

.
end of do-file
```

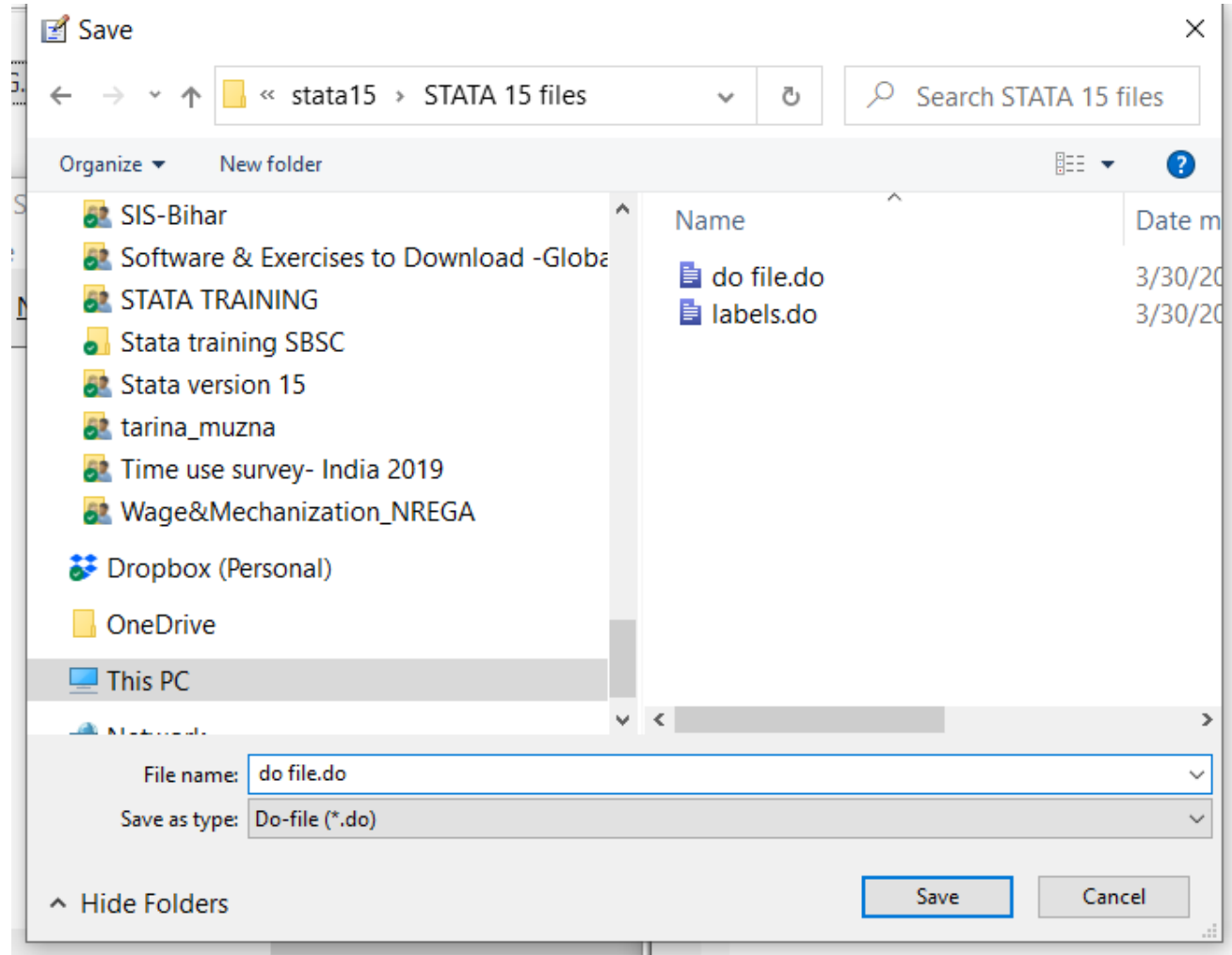
Save a Do file

Click File → Save as..
Give name to DO file
Press Save



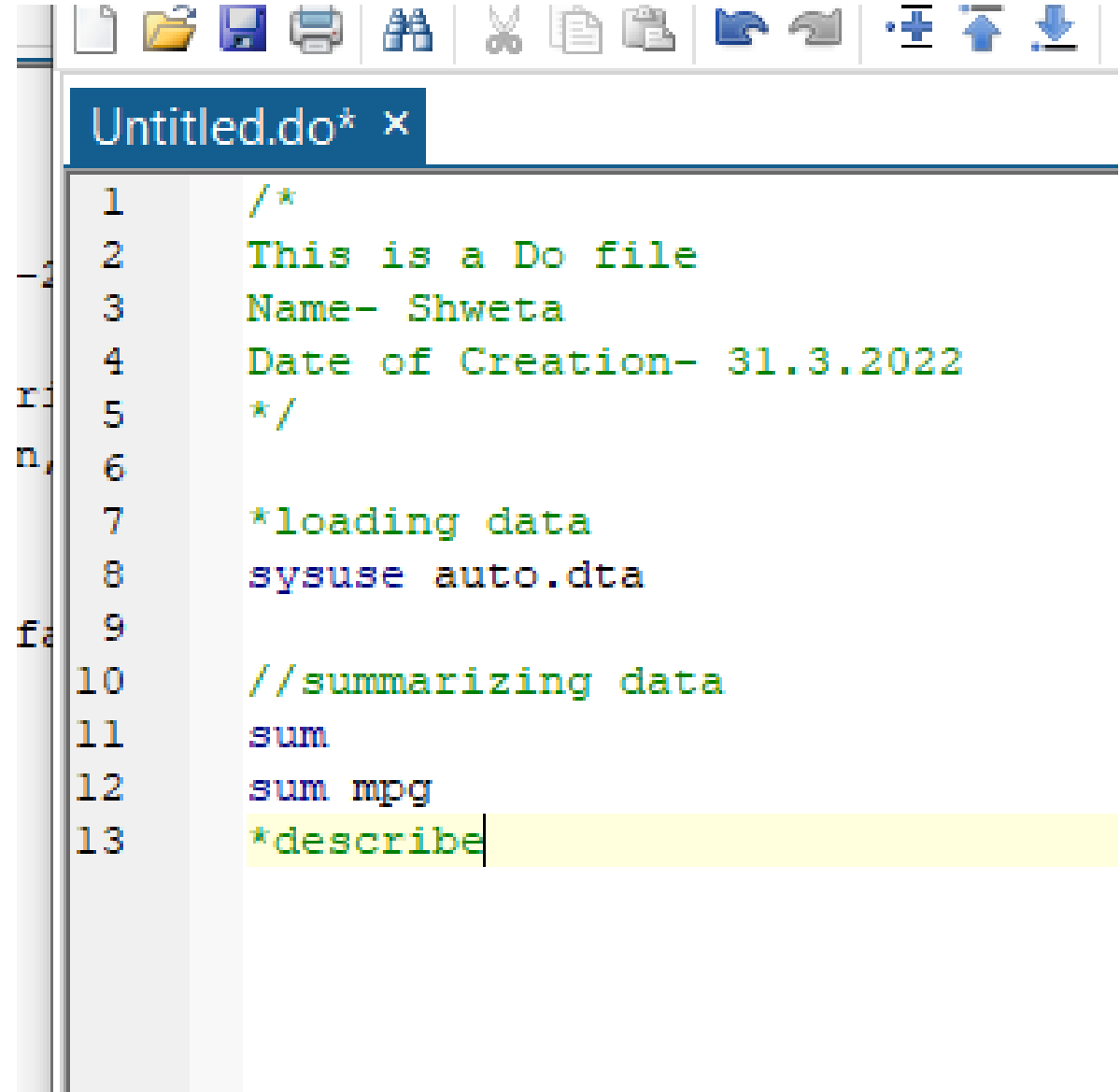
Save a Do file

Click File → Save as..
Give name to DO file
Press Save



Insert comments in Do file

- Comments are those rows that you don't want to execute
- They are used to insert any note, specific comment, or skip a particular line of code
- Insert using
 1. /* */ → anything that comes b/w /* and */ will be not executed
 2. * → a particular line won't be executed
 3. // → same as above



The screenshot shows a text editor window titled "Untitled.do*" with a toolbar at the top. The editor contains the following code:

```
1 /*
2 This is a Do file
3 Name- Shweta
4 Date of Creation- 31.3.2022
5 */
6
7 *loading data
8 sysuse auto.dta
9
10 //summarizing data
11 sum
12 sum mpg
13 *describe
```

The line 13 is highlighted in yellow.

Part 3: Data transformation

The display command

di expression

- Helps to simply display something
- Mostly used to perform quick calculations

```
di 2+2
```

```
di (25*50)/20
```

```
di "mpg"
```

```
di "Sum of 1 and 2=" 1+2
```

```
help di
```

Changing attributes of existing variables *but keeping original data intact*

1. Change name of variable

```
rename oldname newname  
rename mpg mileage
```

2. Change variable label

```
label var varname "write label here"  
label var price "Price in dollars"
```

3. Change value labels

```
label define car_type 0 "Domestic car" 1 "Foreign car"  
label values foreign car_type
```

4. View and save existing labels in directory

```
label list  
label save using "C:\Users\shweta.gupta\Desktop\stata15\STATA 15  
files\labels.do", replace
```

Don't use long names
Only 32 characters allowed
No spaces
New name

Changing attributes of existing variables *but keeping original data intact*

5. Reduce decimals in numeric variable

```
format gear_ratio %6.1f
```

6. Reduce no. of characters in string

```
format make %-10s
```

7. Change justification

```
format make %18s →right-justified
```

```
format price %-8.0gc →left-justified
```

```
format mpg %~8.0g →centre-justified
```

This does not
change data, only
appearance in
browser changes

Changing values of existing variables

this changes the data values

1. Replace values

`replace price=5000` → replaces price by 5000 in all observations

`replace price=5000 if price==5500` → replace price by 5000 only when price=5500

`replace mileage=30 if foreign==0` → replace mpg by 30 in those obs. where foreign=0 or Domestic

`replace mpg=30 if make=="Ford Fiesta"` → replace mpg by 30 in those obs. Where car type (make) is Ford Fiesta

`replace price=price+1000 if make!="Ford Fiesta"`

`replace price`

`Replace trunk= mileage+125`

Changing values of existing variables

this changes the data values

2. Recode values

```
recode 0=1 1=2
```

```
tab foreign, nol
```

```
recode rep78 1=400
```

```
tab rep78
```

- This replaces in foreign, all values that were 0 before to 1, and all values that were 1 before to 2
- Useful when categorical data

Generating new variables

gen newvarname= expression

1. Generating using math operations

```
gen newvar=2000
```

```
gen mpg2= mpg+2
```

```
gen mpg78= mpg+rep78
```

```
gen weight2= weight-trunk
```

```
gen length2= length*2
```

```
gen headroom2= headroom/2
```

```
gen headroom3= headroom/rep78
```

```
gen pricesq= price^2
```

```
gen priceroot= price^(1/2)
```

```
gen lnprice= ln(price)
```

```
gen logprice= log(price)
```

```
gen erep78= exp(rep78) →  $e^{\text{rep78}}$ 
```

Generating new variables

2. Generating using string variables

```
gen make2= "Type"  
gen make3= make2 + ":" + make
```

Note:

- string and numeric cant be combined
- Numeric can be combined with numeric only
- String can be combined with string only
- In string, only the addition of strings is allowed

Generating new variables

3. Generating using **cond** option

Helps create categorical variables

gen newvar= cond(condition,value of newvar if condition is TRUE,value of newvar if cond is FALSE)

gen high_price= cond(price>3000,1,2)

newvar → high_price

Condition → price > 3000

Value of high_price if price > 3000 → 1

Value of high_price if price ≤ 3000 → 2

Condition command

*new var= high_price

*high_price=1000 if price>1000

*high_price=500 if price<=1000

```
gen high_price= 1000 if price>1000
```

```
replace high_price=500 if price<=1000
```

```
gen high_price= cond(price>1000,1000,500)
```

```
gen high_price=cond(price<=1000,500,1000)
```

Generating new variables

4. Generating dummy variables

Dummy variable takes value 0 or 1

Use `gen`, `replace`, `if`

```
gen newvarname=1 if condition=TRUE
```

```
replace newvarname=0 if condition=FALSE
```

```
gen high_price2= 1 if price>3000
```

```
replace high_price2=0 if price<=3000
```

Use the `gen`, `cond` option to create dummy

Generating new variables

5. Generating dummy variables from discrete data

`tab rep78, gen(repdummy)` → gives dummy variable for each category of rep78

repdummy1 =1 if rep78 equals 1
=0 otherwise that is, rep78 =2/3/4/5

repdummy2 =1 if rep78 equals 2
=0 otherwise

repdummy3 =1 if rep78 equals 3
=0 otherwise

repdummy4 =1 if rep78 equals 4
=0 otherwise

repdummy5 =1 if rep78 equals 5
=0 otherwise

Generating new variables

6. Generating specialized variables- egen command

egen newvarname= expression

egen meanrep= mean(rep78) → generates meanrep = mean of rep78 in all observations

egen medrep= median(rep78) → generates medrep = median of rep78 in all observations

egen modrep= mode(rep78) → generates modrep = mode of rep78 in all observations

min, max, range, count, total

Do *help egen* to show options

Generating new variables

7. egen to add variables

Adding variables:

`egen reptr= rowtotal(rep78 trunk)` → generates reptr = sum of rep78 and trunk for each observation

Compare this with gen:

`gen reptr2= rep78 + trunk`

`reptr2=reptr`

gen replaces reptr2 by missing (.) if rep78 or trunk is missing in any obs.

egen treats missing as 0 while calculating the sum

Generating new variables

8. Create unique ID

`gen newvar= _n` → creates ID variable with values 1,2,3...

`gen id= _n`

`bysort rep78: gen id2= _n`

9. Create unique ID by groups

`egen newvar= group(grouping variable)`

`egen id3= group(rep78)`

Headroom has 8 classes, so id2 takes the value 1,2,3,4,5,6,7,8

Generating new variables

10. Create unique ID within groups

`bysort group-varname: gen newvar= _n` → creates ID variable with values 1,2,3...within each subgroup of group-varname

`bysort rep78: gen newid= _n` → creates newid variable=1,2,3... within each category of rep78 (5 categories=1,2,3,4,5)

Generating new variables

11. Get total number of observation

`gen newvar= _N` → creates a new variable = total no. of observations in data

`gen idtot= _N` → takes the value 74 in our data

12. No of observations per sub-group

`bysort rep78: gen sub_idtot= _N` → creates new variable= total obs. per category in rep78 (5 categories=1,2,3,4,5)

`br rep78 newid sub_idtot`

Remove variables or observations

- Keep/drop variables and observations

`keep varlist`

`keep mpg price`

`keep if price==2000`

`keep price*` → keeps all variables with **price** in their name (price, pricesq)

`keep t*` → keeps all variables starting with **t**

`keep *t` → keeps all variables ending with **t**

`keep *ea*` → keeps only those vars which have **ea** in their name

`drop varlist`

`drop mpg`

`drop if price==2000`

- Be careful while dropping
- Data can't be restored once you drop
- Best practice- save the data before working on it

Changing variable type

- Convert numeric variable to string
`tostring oldvarname, gen(newvarname)`
`tostring price, gen(price2)`

```
tostring varname(s), replace  
tostring price mpg, replace
```

- Convert string to numeric

This is possible only when there are numbers stored as string

```
destring oldvarname, gen(newvarname)  
destring price2, gen(price3)
```

```
destring oldvarname, replace  
destring price2, replace
```

Installing commands in STATA

```
ssc install commandname
```

```
findit commandname
```

```
help commandname
```

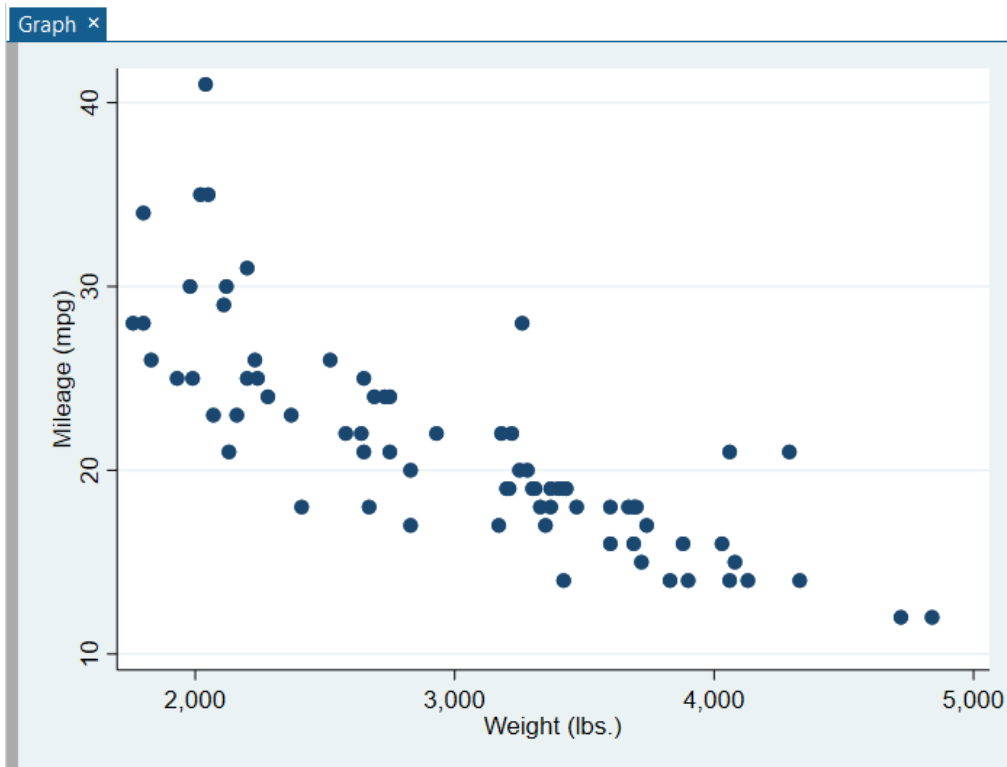
Part 4: Data visualization

scatterplot

Two variable plot-

scatter y x

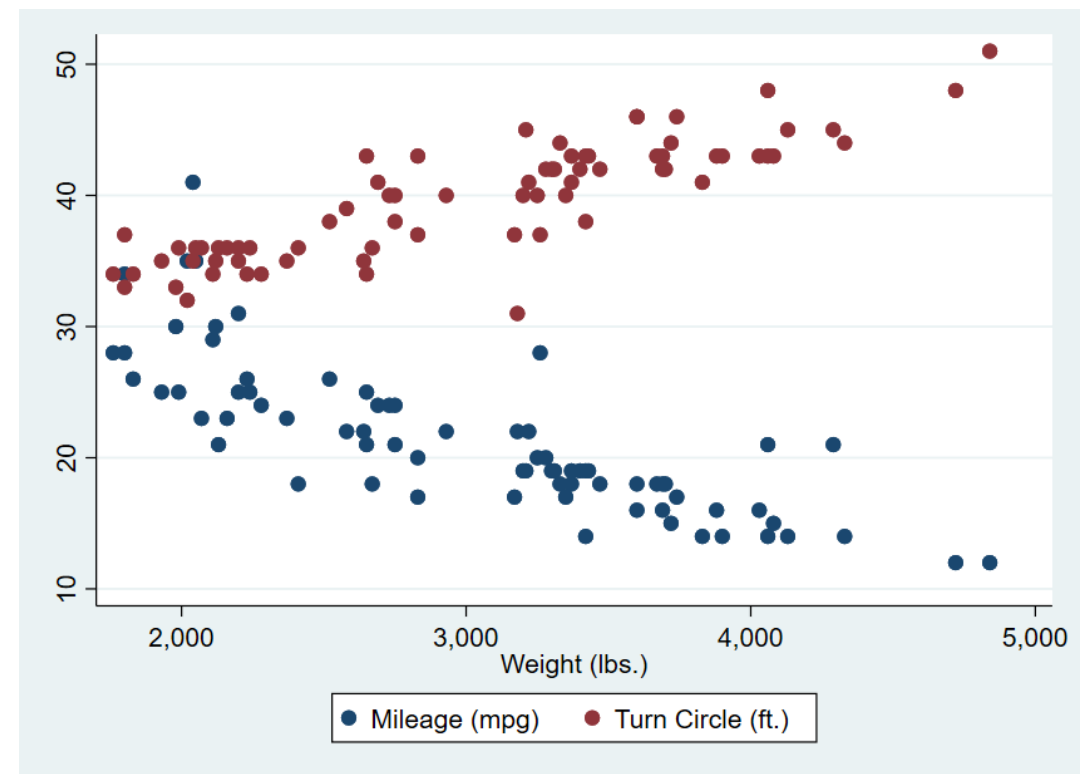
scatter mpg weight



Multiple variables-

scatter $y1$ $y2$ x

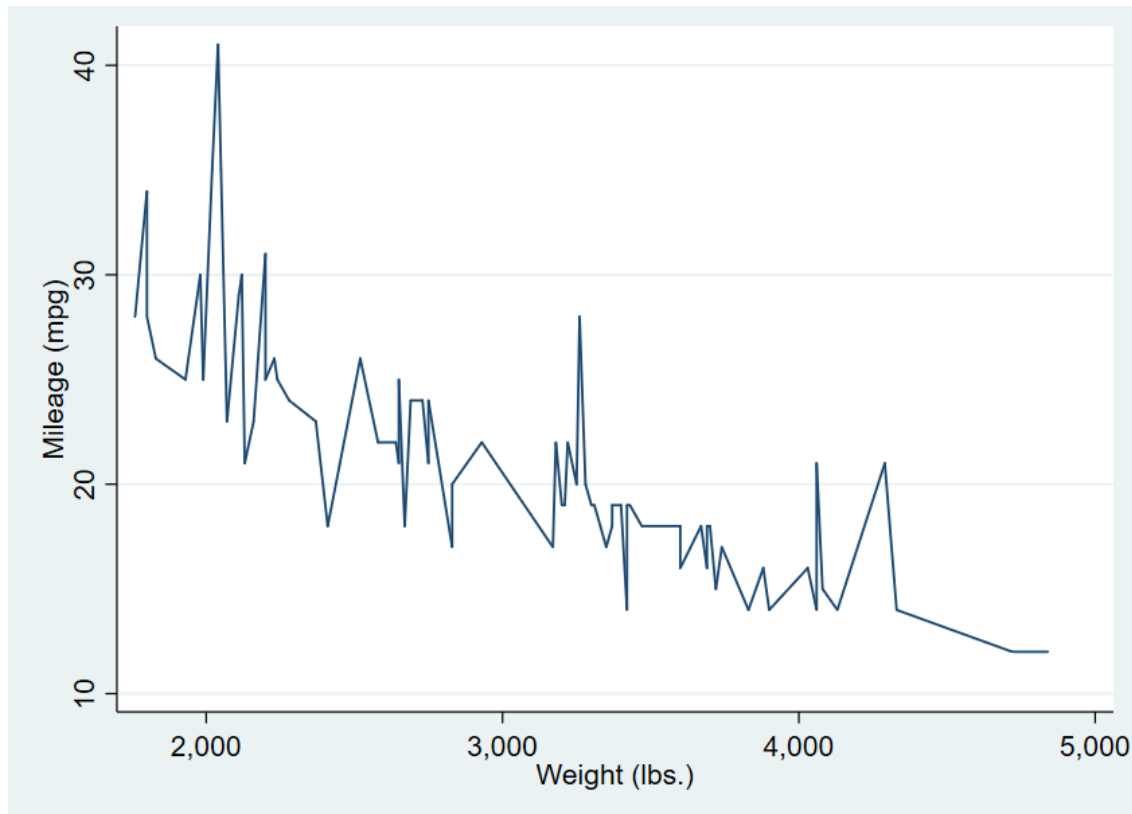
scatter mpg turn weight



Line graphs

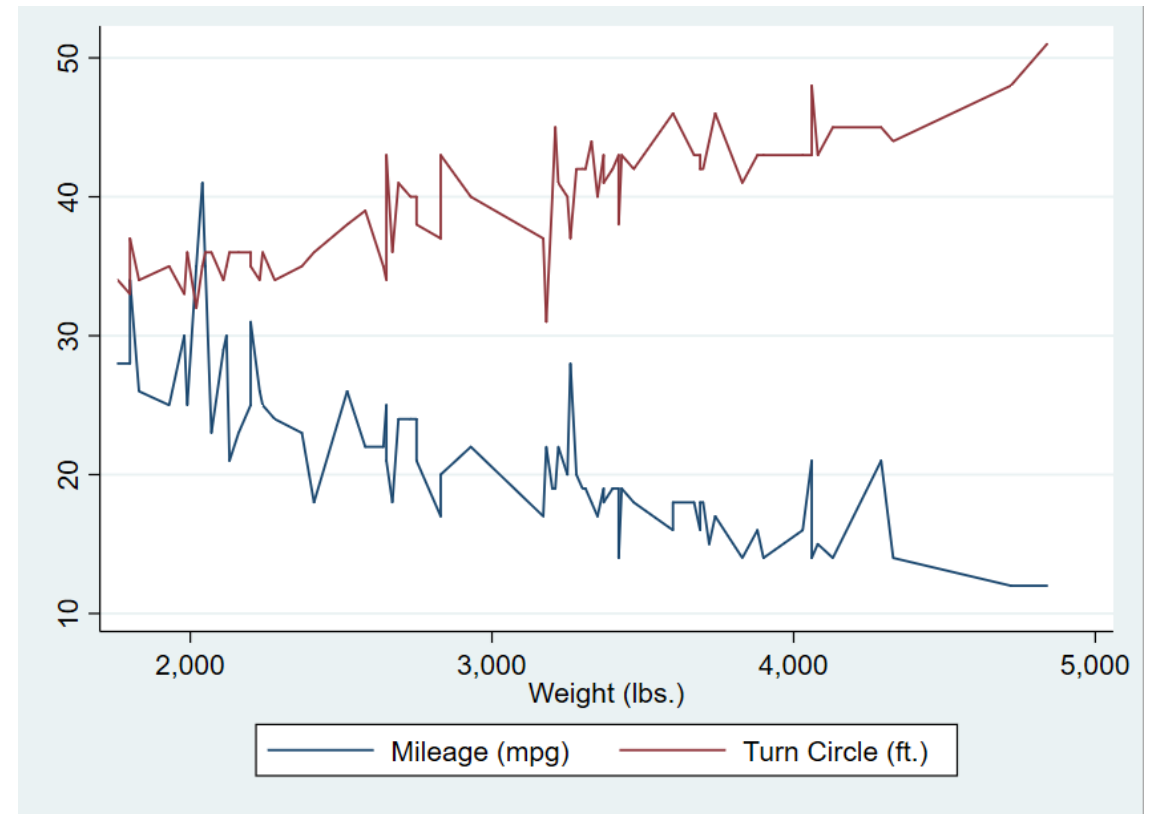
Two variable plot-

twoway line y x, sort
twoway line mpg weight



Multiple variables-

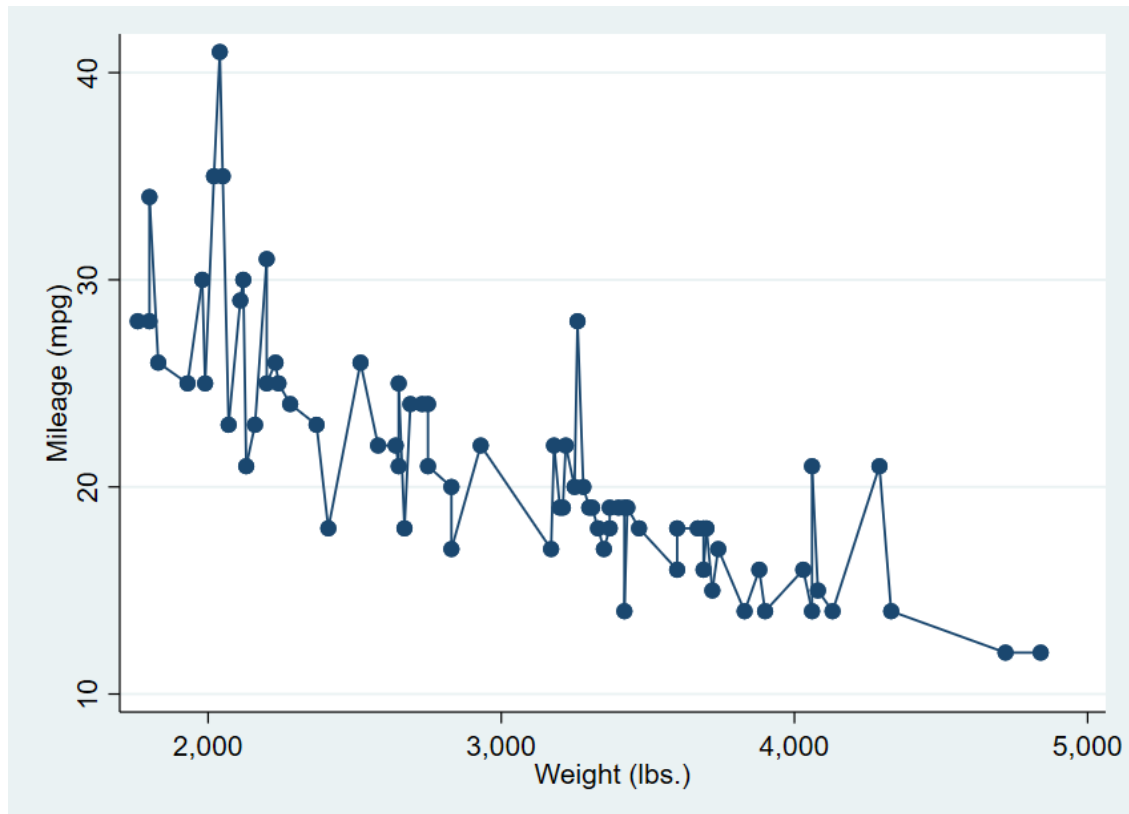
twoway line y1 y2 x, sort
twoway line mpg turn weight, sort



Scatter+Line graphs

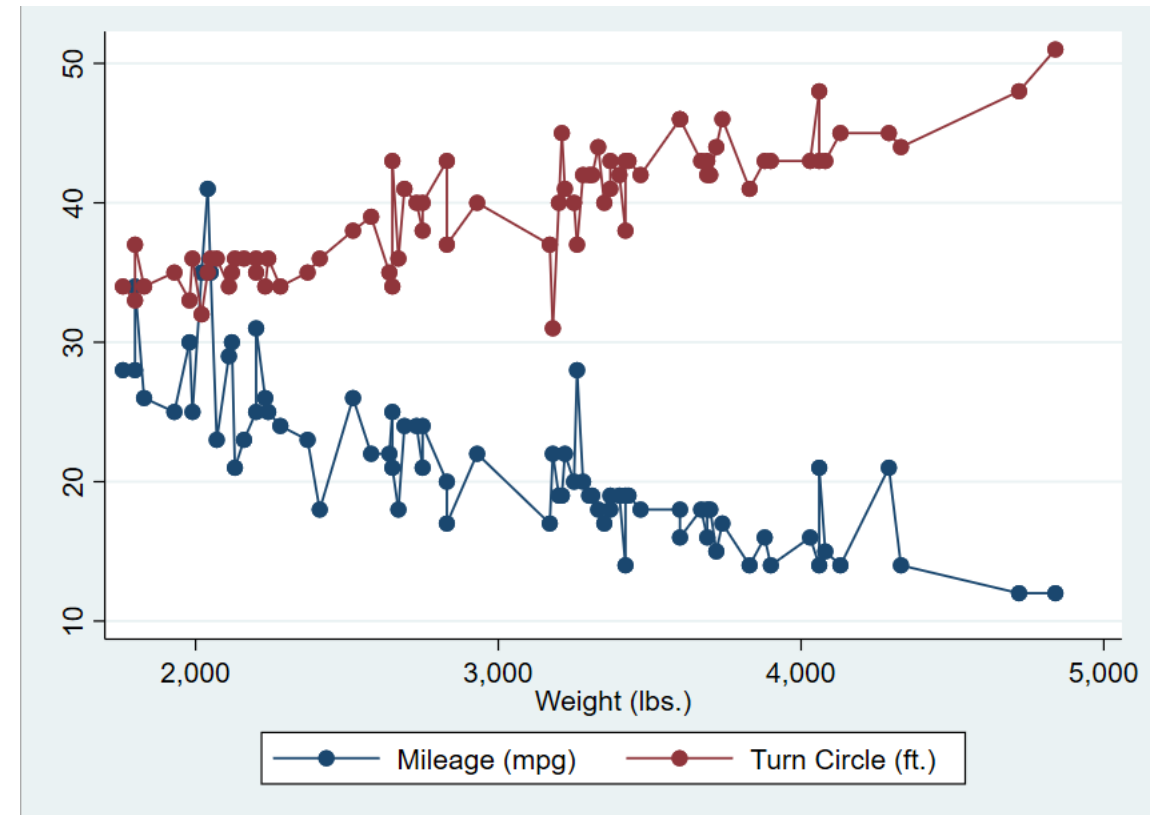
Two variable plot-

twoway connected y x, sort
twoway connected mpg weight



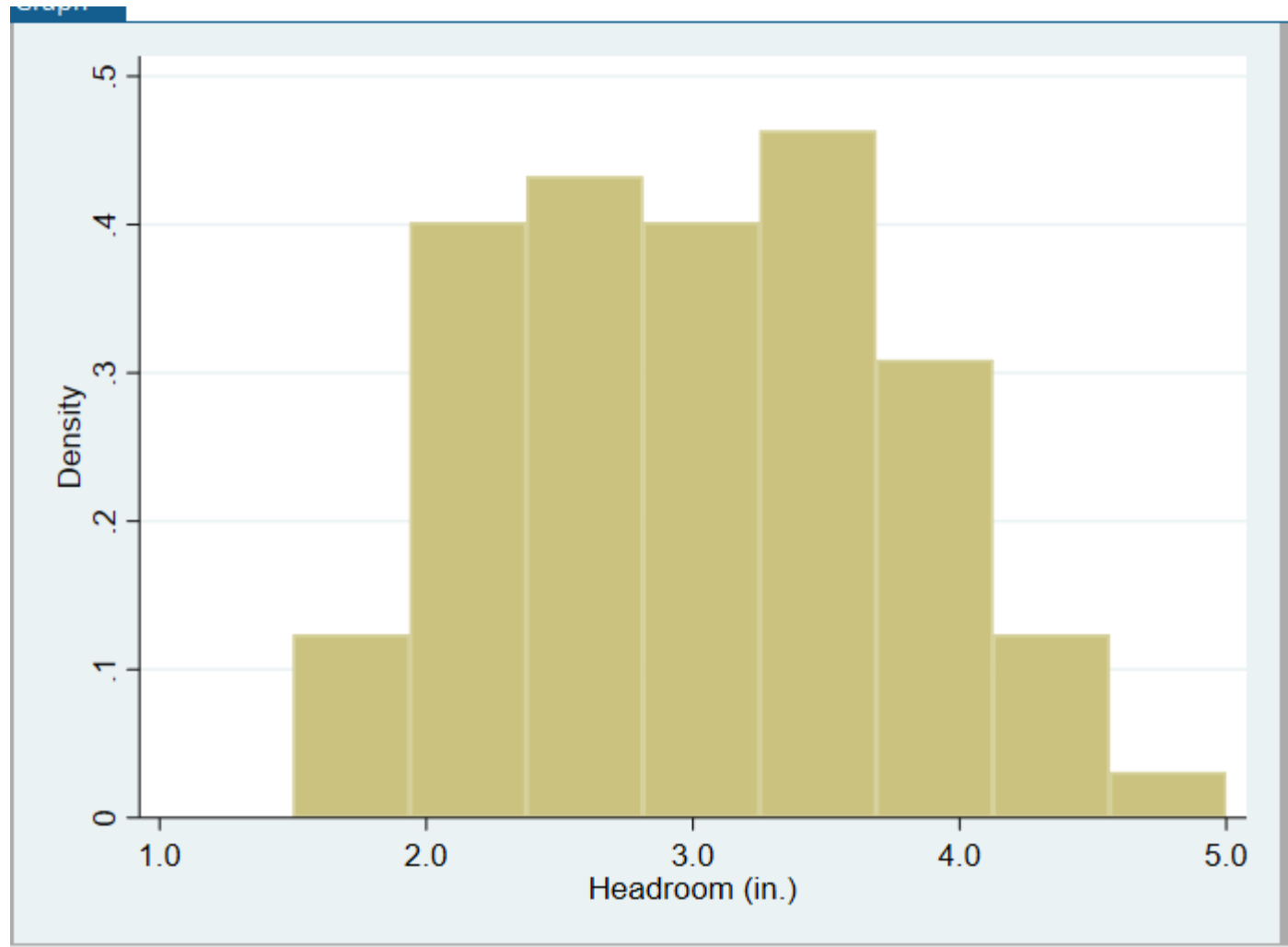
Multiple variables-

twoway connected y1 y2 x, sort
twoway connected mpg turn weight, sort




Histogram

- Use the toolbar for best results
 - histogram *x, options*
 - histogram headroom →
 - hist head, freq
 - hist head, percent

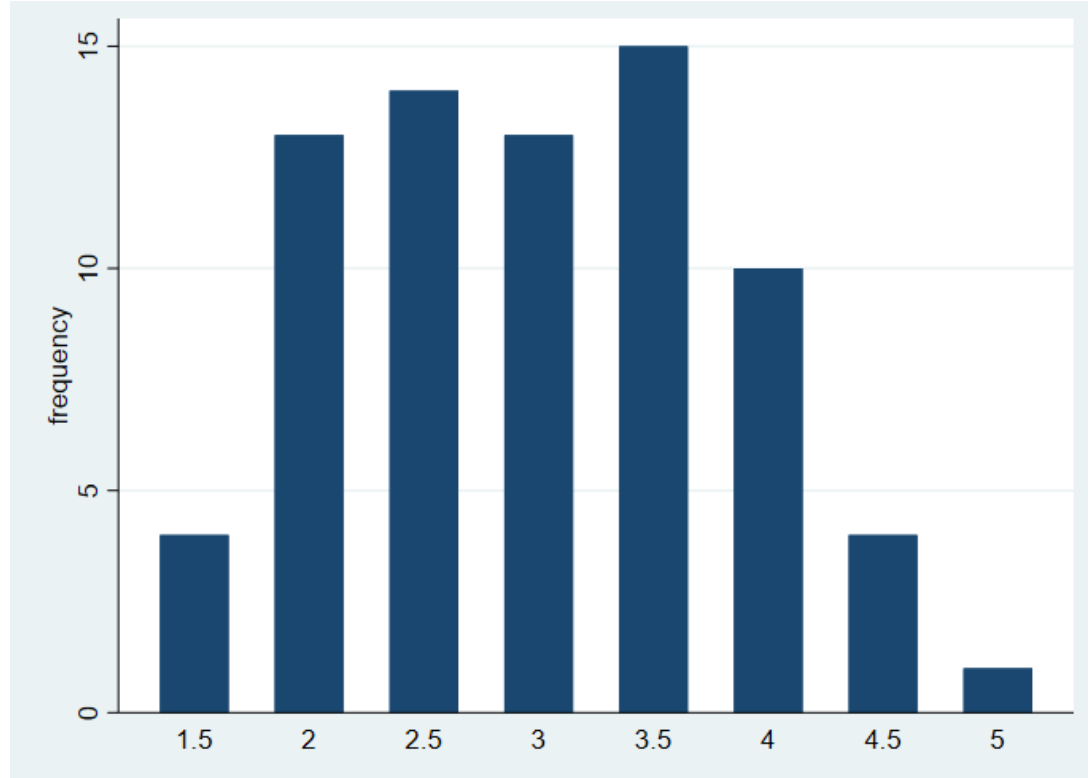
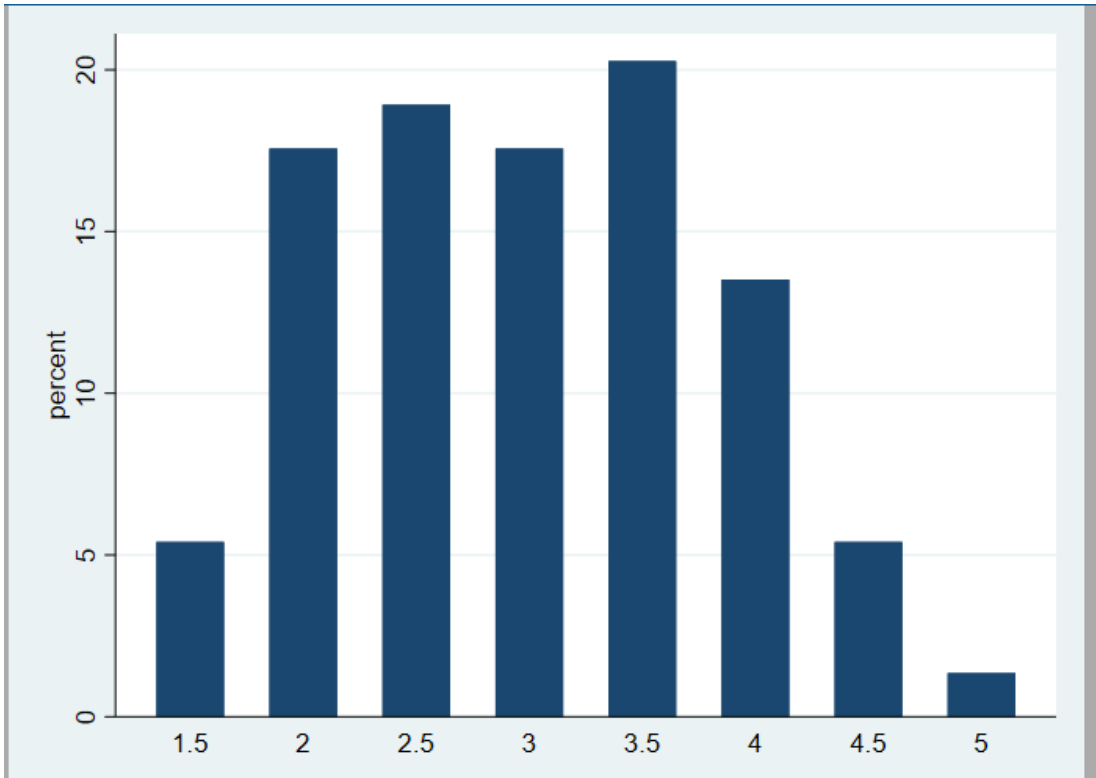


Bar graph

Percent by default

graph bar, over(x) 
graph bar, over(head)

graph bar (count), over(x)
graph bar (count), over(head)

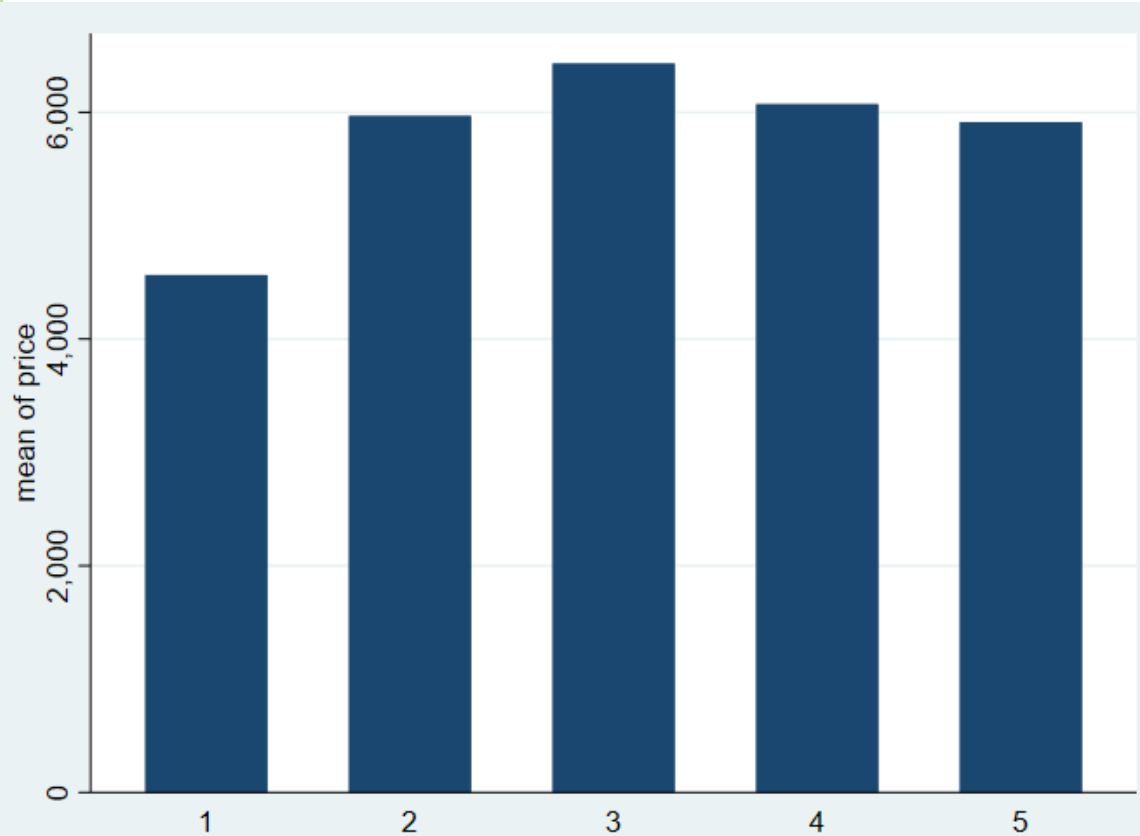


Bar graph

```
graph bar (mean) y, over(x)
```

```
graph bar (mean) price, over(rep78)
```

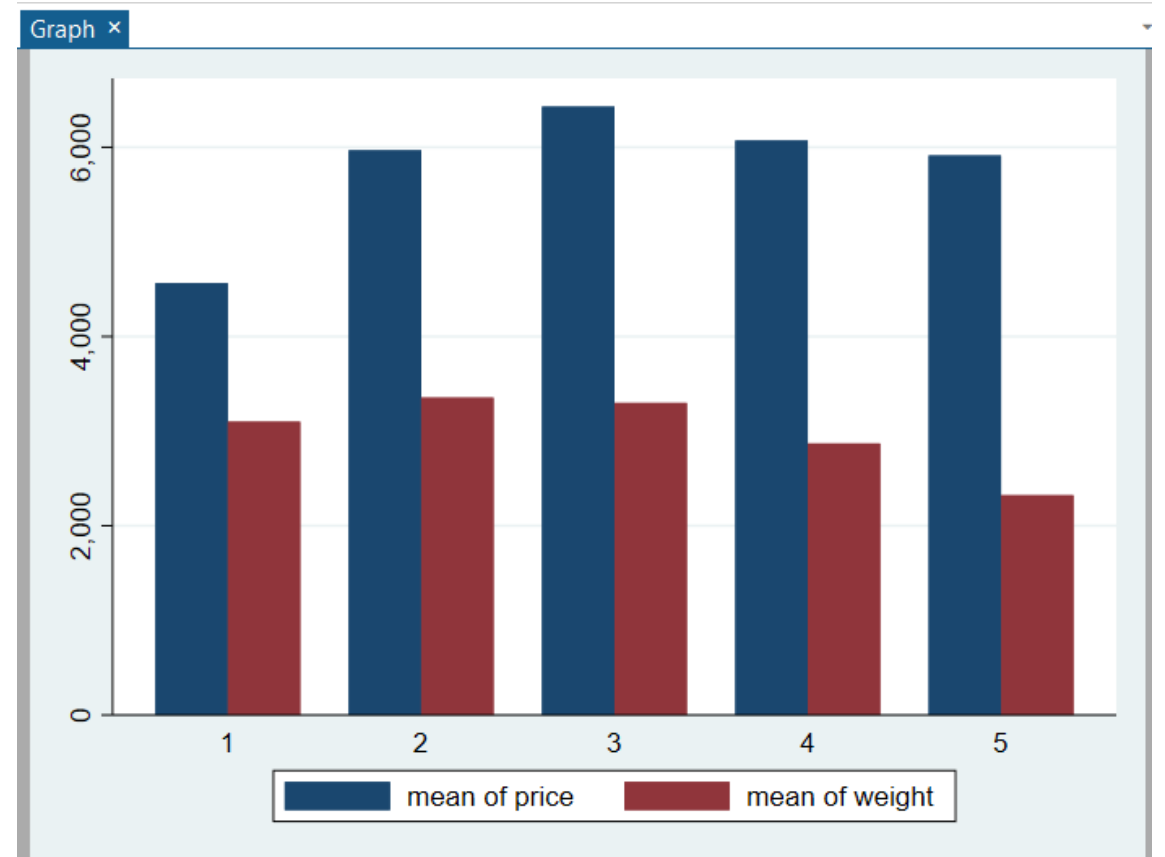
```
tab rep78, sum(price)
```



```
graph bar (mean) y1 y2 , over(x)
```

```
graph bar (mean) price weight, over(rep78)
```

```
table rep78, c(mean price mean weight)
```



Density curve

`kdensity x`

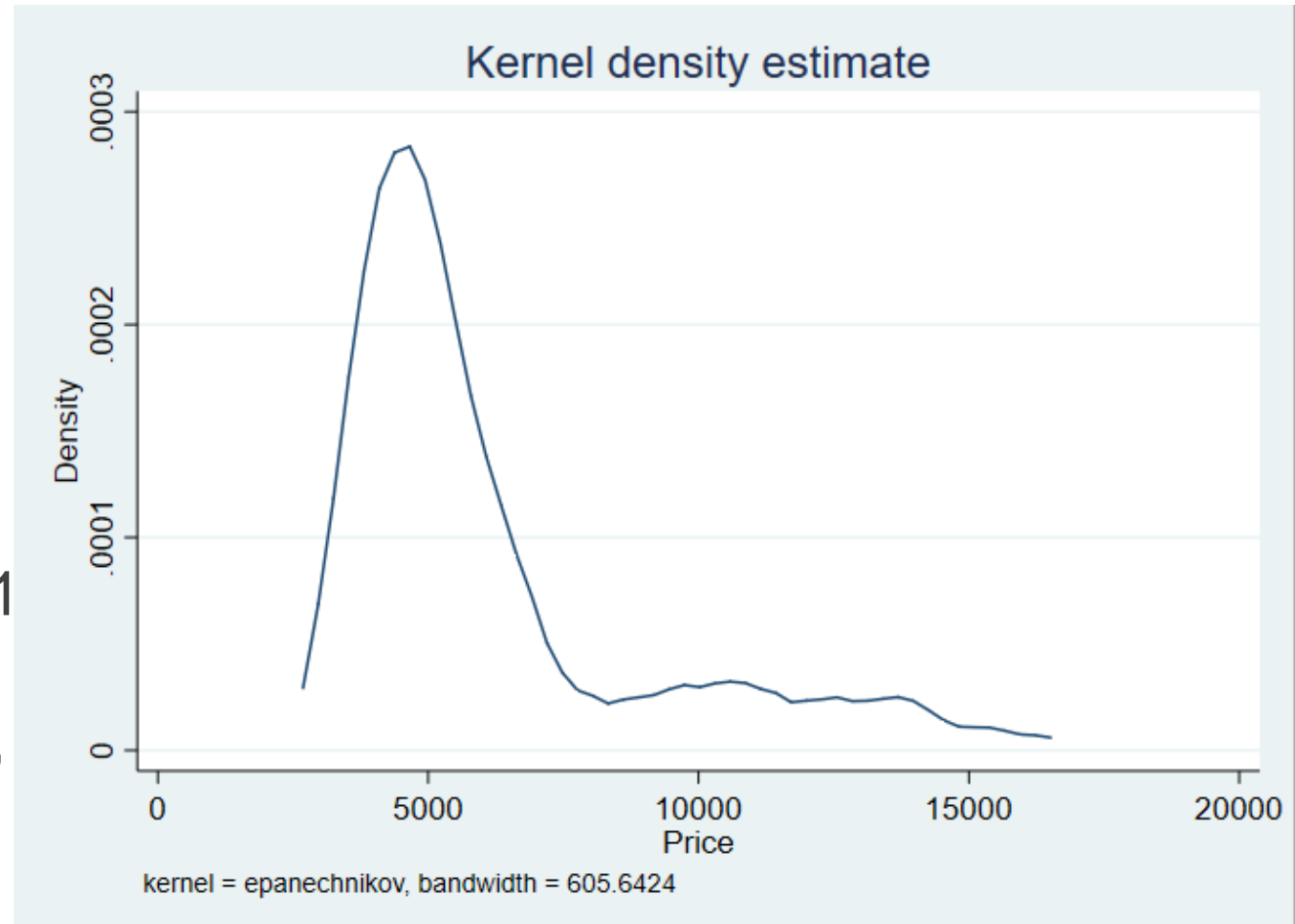
`kdensity price`

- Gives the distribution of the variable-

- area on y-axis
- Range of values of the variable on x-axis

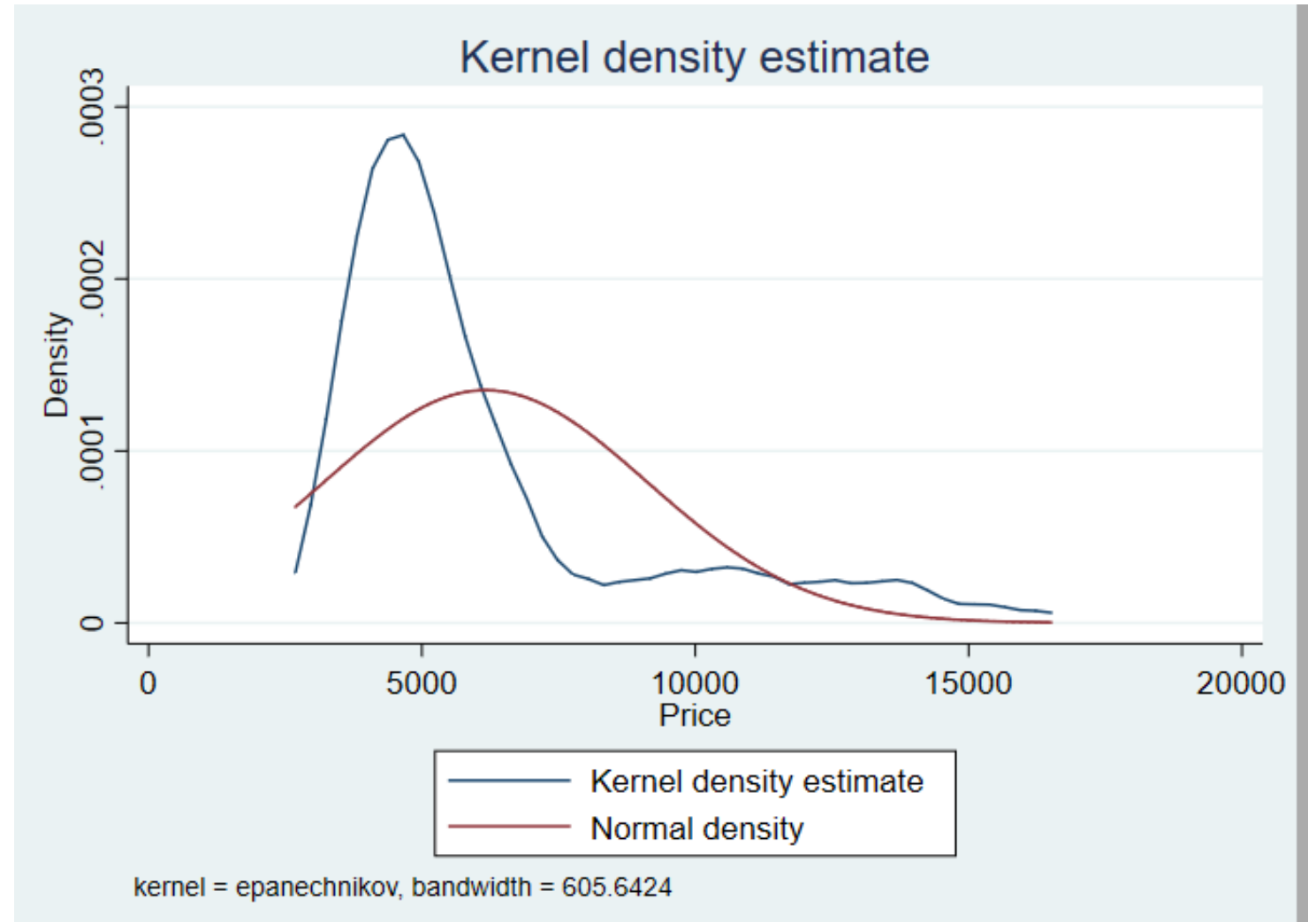
NOTE: Area under the curve==1
ALWAYS

- Can plot only 1 variable at a time, but can be combined using `twoway`



Density curve

kdensity x, normal
kdensity price, normal



Graph attributes

1. Labelling axis-

```
scatter mpg weight, ytitle(Mileage of car) xtitle(Weight of car)
```

2. Title, subtitle of main graph-

```
scatter mpg weight, title(Scatter Plot)
```

3. Background color

```
scatter mpg weight, graphregion(color(green))
```

4. Colour scheme of entire graph

```
scatter mpg weight, scheme(economist)
```

5. Scale of graph

```
scatter mpg weight, xsize(10) ysize(10)
```


Graph attributes

6. Assign name to graph in directory

```
scatter mpg weight, name(scatter1, replace)
```

```
scatter mpg turn weight, name(scatter2, replace)
```

Graph attributes

6. Saving a graph- editable in STATA format

`graph save using "C:\path\scatterplot.gph", replace`

- Show saving directly using toolbar

7. Exporting a graph- png/pdf/jpg

`graph export using using "C:\path\scatterplot.png", replace`

- Show saving using toolbar

Graph using IF condition

`scatter mpg weight if price>2000`

- Plots pairs of mpg and weight for which `price>2000`

`scatter mpg weight if foreign==1`

- Plots pairs of mpg and weight for which `foreign=1`

Plotting multiple graphs together

- Overlaying graphs

```
scatter mpg weight || scatter turn weight  
tw (scatter mpg weight ) (scatter turn weight)  
scatter mpg turn weight
```

- Using if condition

```
scatter mpg weight if foreign==0 || scatter mpg weight if foreign==1
```

- Using *by* in the graph command

```
scatter mpg weight, by(foreign)
```

Plotting multiple graphs together

- Graph combine

```
scatter mpg weight if foreign==0, name(scatter1, replace)
```

```
scatter mpg weight if foreign==1, name(scatter2, replace)
```

```
graph combine scatter1 scatter2
```

```
graph combine scatter1 scatter2, ycommon
```

Plotting graphs conditioned on data

- Combining histogram and frequency distribution
- Generate a variable, say total, with frequency for each class of x variable

```
bysort headroom: gen headtotal= _N
```

```
tw histogram headroom, frequency || connected headtotal headroom
```

Other graphs

- Piechart- graph pie..
- Dot chart- graph dot..
- Box plots- graph box x

What is a `good' graph?

- Easy to understand (not too many lines or points, not clumsy)
- Should deliver the message it is supposed to give
- Always label axes
- Always give a key when plotting multiple variables
- Always give title
- Don't plot too many variables in one graph as much as possible

Question

- Can you plot a graph with 1 or more string variables?

Generating a variable with specific distributions

11. Random variable with normal distribution

`gen newvar= rnormal(m,s)` → creates a new variable that follows normal distribution with mean=m, SD=s

`newvar ~N(m,s)`

Normality is a law of large samples

a) When you don't have any data in your directory

`clear all`

`set obs 10`

`set seed 15678`

`gen normalvar= rnormal(4,10)`

`sum normalvar`

`kdensity normalvar`

Write this in a Do file.

Observe how summary and kdensity changes when N=10, 100, 1000, 10000, 10000: Dist. Becomes closer to bell curve with mean=4, SD=10.

Generating a variable with specific distributions

11. Random variable with normal distribution

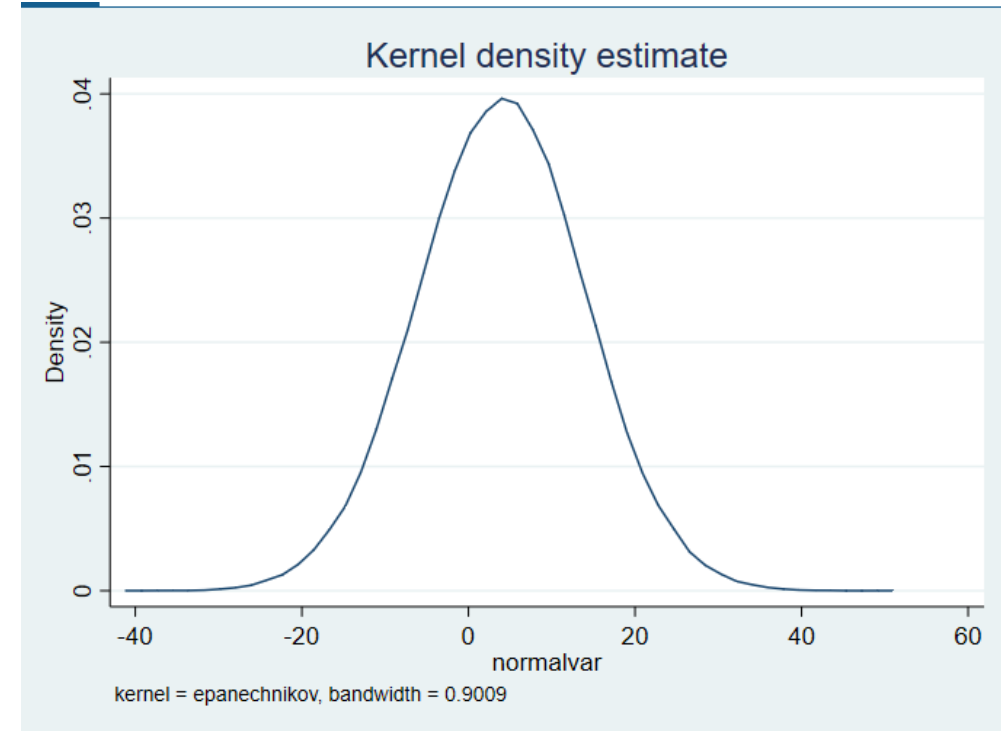
b) If some data already exists in your directory

```
gen normalvar= rnormal(4,10)
```

```
sum normalvar
```

```
kdensity normalvar
```

Setting seed helps to get the same set of random numbers every time you run the code



Generating a variable with specific distributions

12. Random variable with standard normal distribution

```
gen newvar= rnormal(0,1)
```

```
newvar ~Z(m,s)
```

```
gen newvar= rnormal()
```

Generating a variable with specific distributions

12. Uniform distribution

```
gen newvar= runiform(a,b)
```

where,

a= minimum and b= maximum

Mean=(a+b)/2

Variance=(b-a)²/12

```
clear all
```

```
set obs 10000
```

```
set seed 15678
```

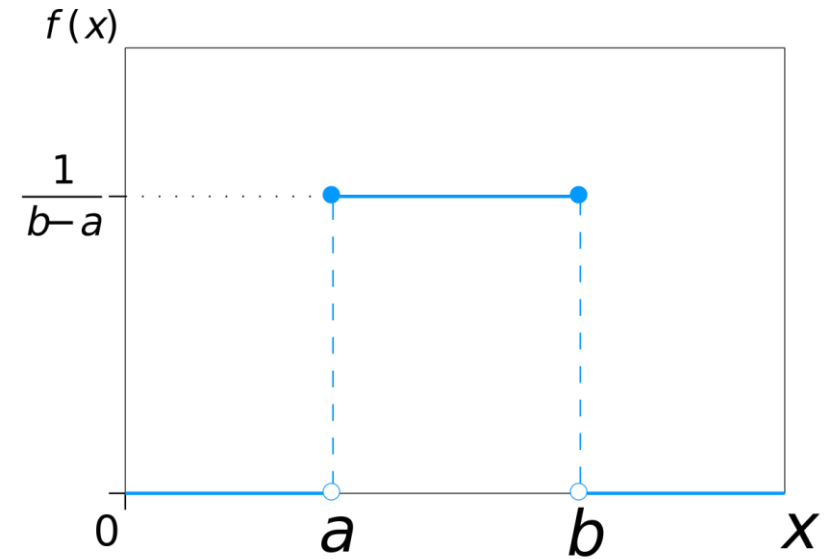
```
gen uniformvar= runiform(4,10)
```

```
sum uniformvar
```

```
kdensity uniformvar
```

Now find mean and variance from sum and check

```
help random
```



Working with multiple datasets

- Merge 2 data files
- Append 2 or more data files

Advanced: Changing layout of data

- Reshape
- Wide to long
- Long to wide