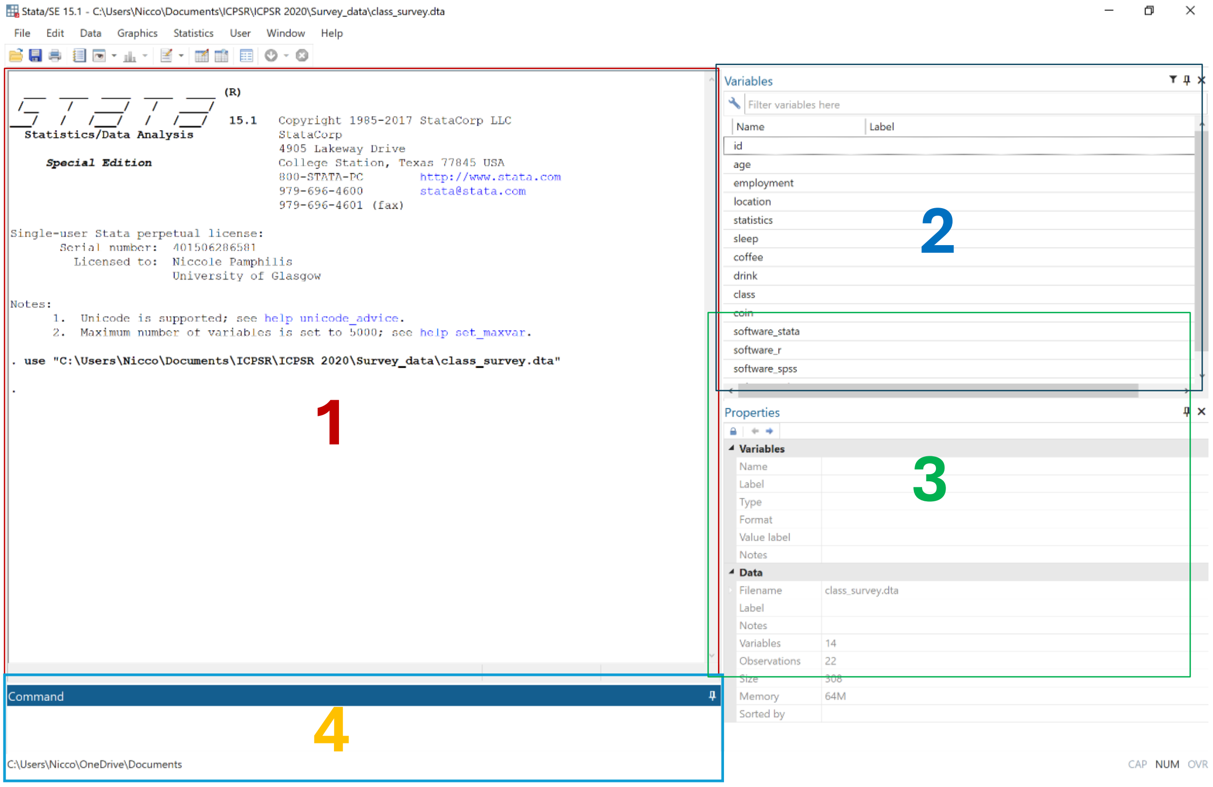
Introduction to STATA
Dr. Niccole M. Pamphilis

Stata interface, Entering data, Do files, Basic Commands

Welcome to this session on Stata. The objective for this session is to become a bit more familiar with Stata. We will be covering what the Stata interface looks like, how to open data, how to create a data set in Stata, and a few very basic commands. This session is not designed to explore model selection or the interpretation of output.

Stata Interface:



```
 ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  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:  401506286581
         Licensed to:  Niccole Pamphilis
                       University of Glasgow

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "C:\Users\Nicco\Documents\ICPSR\ICPSR 2020\Survey_data\class_survey.dta"

.
```

Name	Label
id	
age	
employment	
location	
statistics	
sleep	
coffee	
drink	
class	
coin	
software_stata	
software_r	
software_spss	

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Filename	class_survey.dta
Label	
Notes	
Variables	14
Observations	22
Size	308
Memory	64M
Sorted by	

C:\Users\Nicco\OneDrive\Documents

1: Output window. This is where the output for your commands will appear.

2: Variable window. This is where you will be able to see the variables available in your dataset.

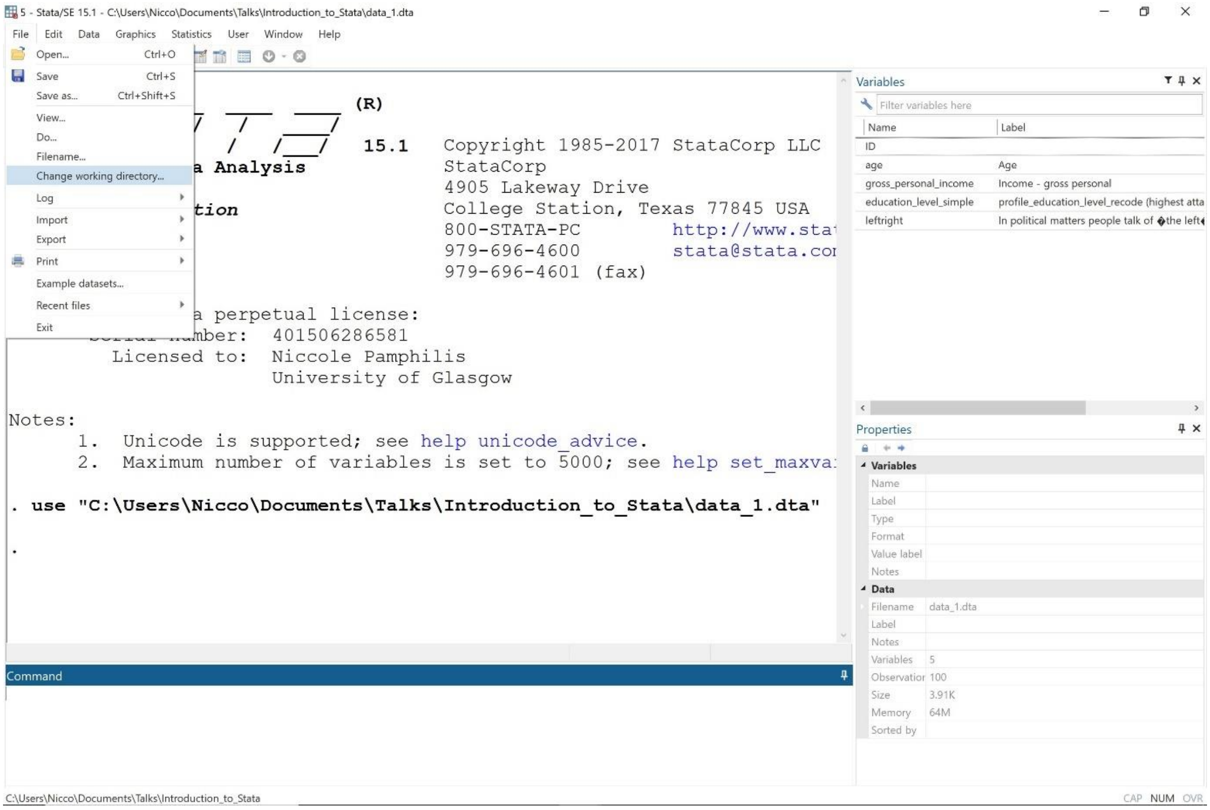
3: Properties window. This is where you will be able to see and change information related to the individual variables being used in a session (listed in the variable window (2)).

4: Command window. This is where you can run your commands/code from.

Working Directory

The working directory shows Stata where to save files from your current coding session. It is useful to create separate files for each project to keep the different results in easy to locate areas.

To set the working directory you can use the drop-down menus:



```
                                          (R)
                 /     /   ____/
                /     /   /___/    15.1   Copyright 1985-2017 StataCorp LLC
                a Analysis                StataCorp
                                          4905 Lakeway Drive
                tion                      College Station, Texas 77845 USA
                                          800-STATA-PC        http://www.sta
                                          979-696-4600        stata@stata.co
                                          979-696-4601 (fax)

                a perpetual license:
              mber:   401506286581
         Licensed to:   Niccole Pamphilis
                        University of Glasgow

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxva

. use "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\data_1.dta"

.
```

Variables

Filter variables here

Name	Label
ID	
age	Age
gross_personal_income	Income - gross personal
education_level_simple	profile_education_level_recode (highest atta
leftright	In political matters people talk of �the left�

Properties

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Filename	data_1.dta
Label	
Notes	
Variables	5
Observatior	100
Size	3.91K
Memory	64M
Sorted by	

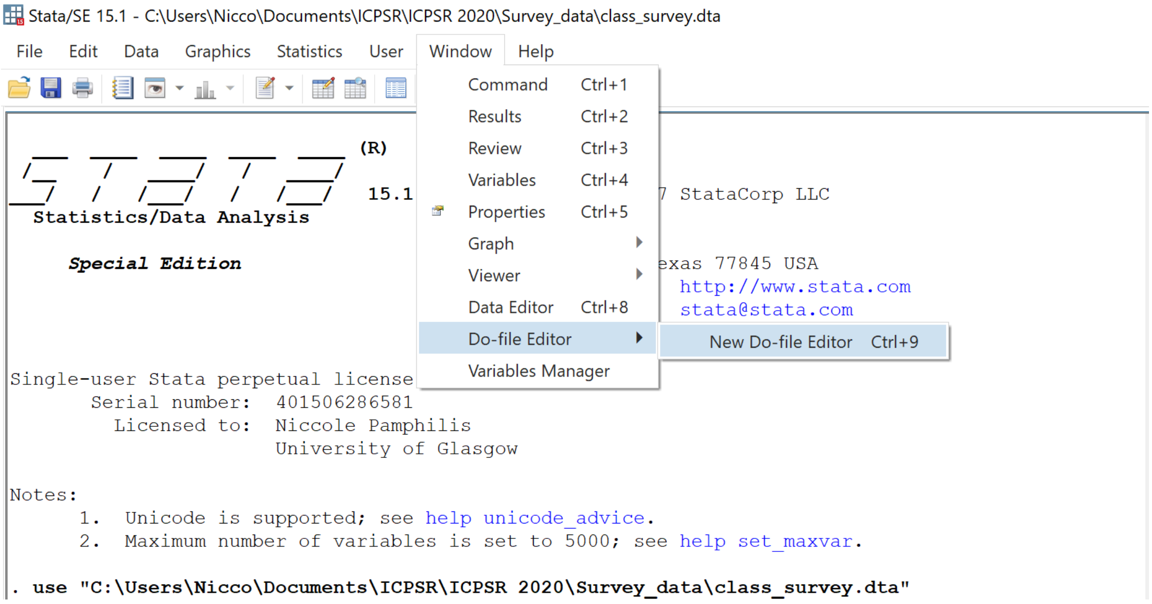
C:\Users\Nicco\Documents\Talks\Introduction_to_Stata

Or you can use the command "cd" for change directory and then indicate the pathway you want to use. Note that Stata uses backwards slashes for pathways.

```
cd "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata"
```

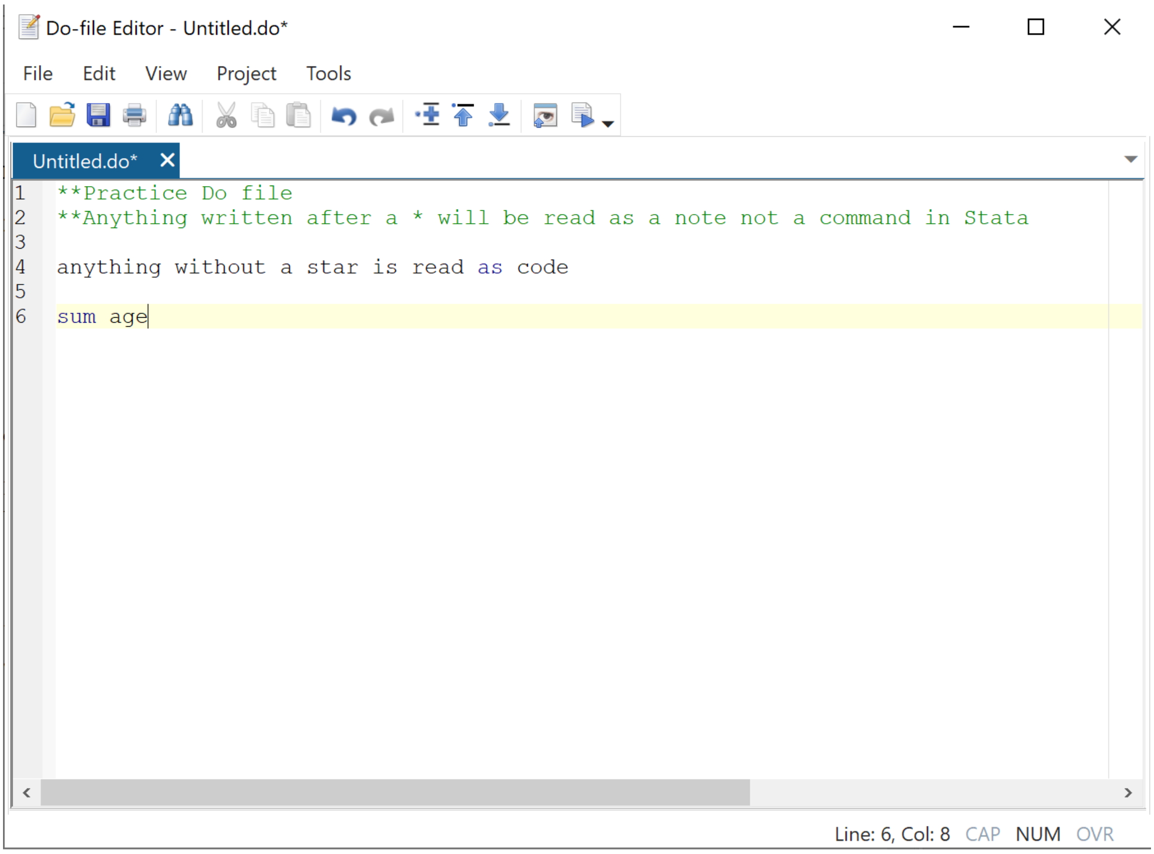
Do files

You can run a single line of code at a time, but a lot of projects cannot be finished in one session and require multiple sessions and corrections. Working with a do.file in Stata allows you to write your code, save your code, and run your code from one file that works with Stata.



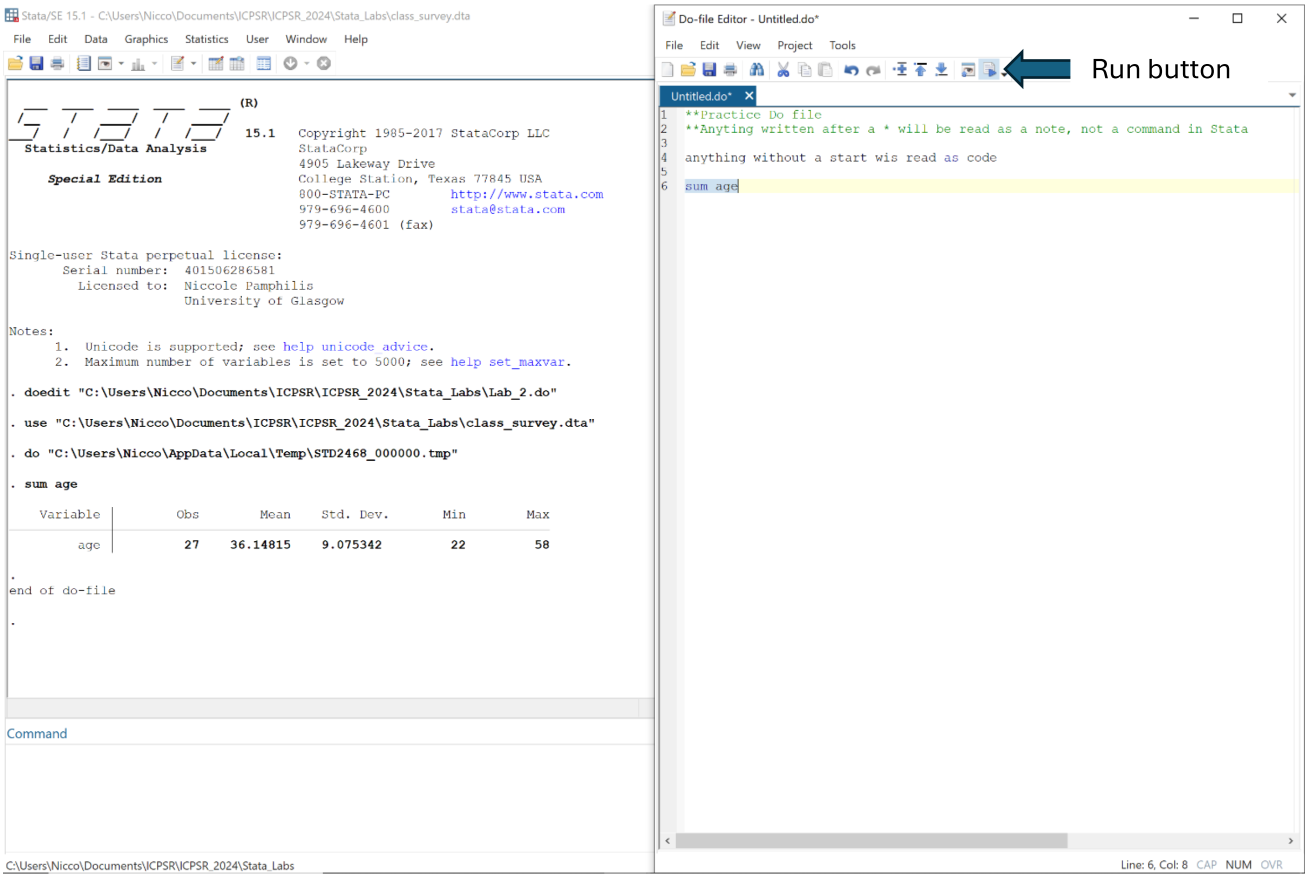
```
                                                      (R)
 Statistics/Data Analysis                         15.1          7 StataCorp LLC

     Special Edition                                            exas 77845 USA
                                                                  http://www.stata.com
                                                                  stata@stata.com

Single-user Stata perpetual license
       Serial number:  401506286581
         Licensed to:  Niccole Pamphilis
                       University of Glasgow

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "C:\Users\Nicco\Documents\ICPSR\ICPSR 2020\Survey_data\class_survey.dta"
```

Do files allow you to make notes in them as well to remind yourself what you are doing and why. Use * at the start of a line that is your personal note, so Stata does not think it is code to run.



```
**Practice Do file
**Anything written after a * will be read as a note not a command in Stata

anything without a star is read as code

sum age
```

You can run commands directly from your do files by highlighting the code you want and pressing the run button.


Stata/SE 15.1 - C:\Users\Nicco\Documents\ICPSR\ICPSR_2024\Stata_Labs\class_survey.dta
File Edit Data Graphics Statistics User Window Help
Command
C:\Users\Nicco\Documents\ICPSR\ICPSR_2024\Stata_Labs
Do-file Editor - Untitled.do*
File Edit View Project Tools
Run button
Untitled.do*
Line: 6, Col: 8 CAP NUM OVR

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  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:  401506286581
         Licensed to:  Niccole Pamphilis
                       University of Glasgow

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. doedit "C:\Users\Nicco\Documents\ICPSR\ICPSR_2024\Stata_Labs\Lab_2.do"

. use "C:\Users\Nicco\Documents\ICPSR\ICPSR_2024\Stata_Labs\class_survey.dta"

. do "C:\Users\Nicco\AppData\Local\Temp\STD2468_000000.tmp"

. sum age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |         27    36.14815    9.075342         22         58

.
end of do-file

.
```

```
**Practice Do file
**Anyting written after a * will be read as a note, not a command in Stata

anything without a start wis read as code

sum age
```

Notes on Do files

Everyone will have their own coding practices, but a few tips for beginners:

1. Annotate your do files, what code are your running, why are you running it, what decisions did you make/why

2. Have a clean do file for your final work separate from the one you used during your analysis stage. What code produced the results in your chapter/paper. This might include annotations such as ##Code used for Table 1, ##Code used for Figure 2).

3. Save separate do files for data set cleaning and the analysis of your data.

Commands in Stata

For those less familiar with coding Stata has two ways commands can be run:

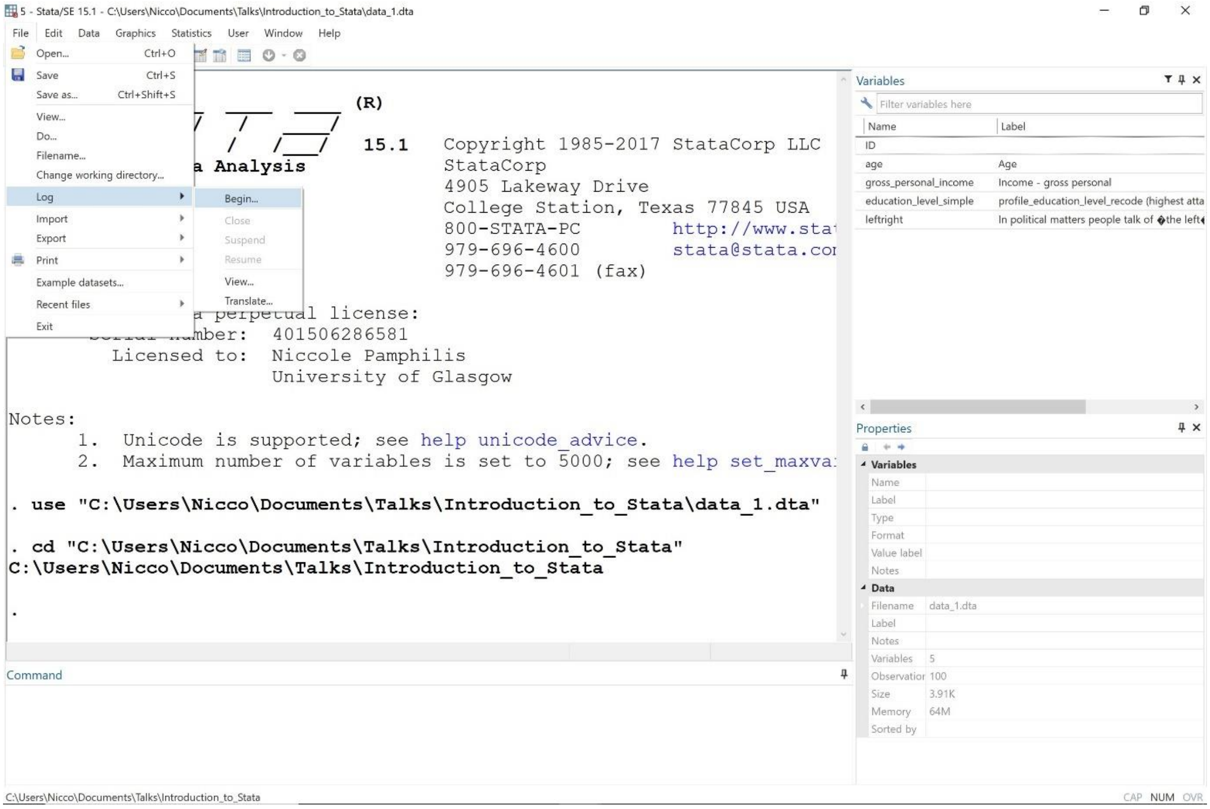
1. Using code entered and run through the command window

2. Using drop down menus, which then show the code used for the commands in the output window.

Log Files

Some researchers prefer to save a copy of their Stata Sessions for reference later. There are referred to as log files. The output window itself only keeps a record of the last 100 or so lines code and output that were run in a session. If you want to look further back at the various models and output you ran during a session a log file might be a useful tool for you to use.

To start/end a log file for a Stata session you can use the drop-down menus:



```
                                       (R)
          /   /    /___/   15.1   Copyright 1985-2017 StataCorp LLC
        a Analysis                StataCorp
                                  4905 Lakeway Drive
                                  College Station, Texas 77845 USA
                                  800-STATA-PC        http://www.stat
                                  979-696-4600        stata@stata.con
                                  979-696-4601 (fax)

        a perpetual license:
       Serial number:  401506286581
         Licensed to:  Niccole Pamphilis
                       University of Glasgow

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxva:

. use "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\data_1.dta"

. cd "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata"
C:\Users\Nicco\Documents\Talks\Introduction_to_Stata

.
```

Name	Label
ID	
age	Age
gross_personal_income	Income - gross personal
education_level_simple	profile_education_level_recode (highest atta
leftright	In political matters people talk of �the left�

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Filename	data_1.dta
Label	
Notes	
Variables	5
Observatior	100
Size	3.91K
Memory	64M
Sorted by	

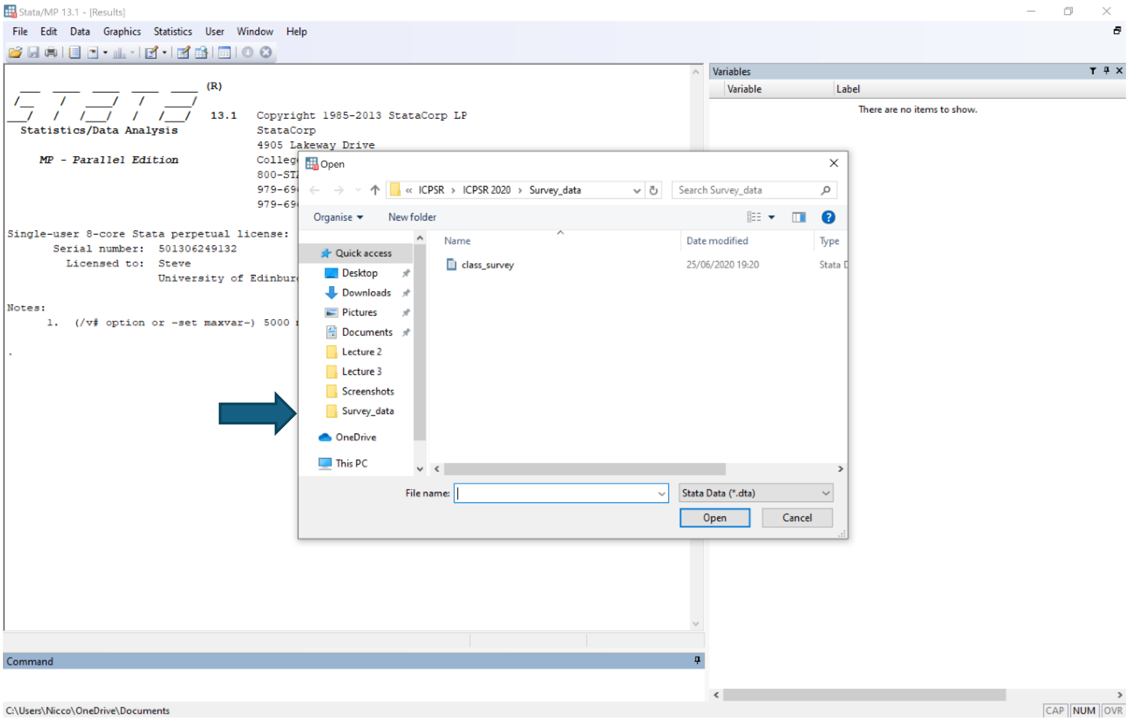
Or you can use the command "log using" and then indicate the file pathway were you would like the log file to be stored:

log using "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\log_file.smcl"

Note that the log file is stored as a Stata file type, so unless you have access to Stata on the workstation you are using, you cannot open this file.

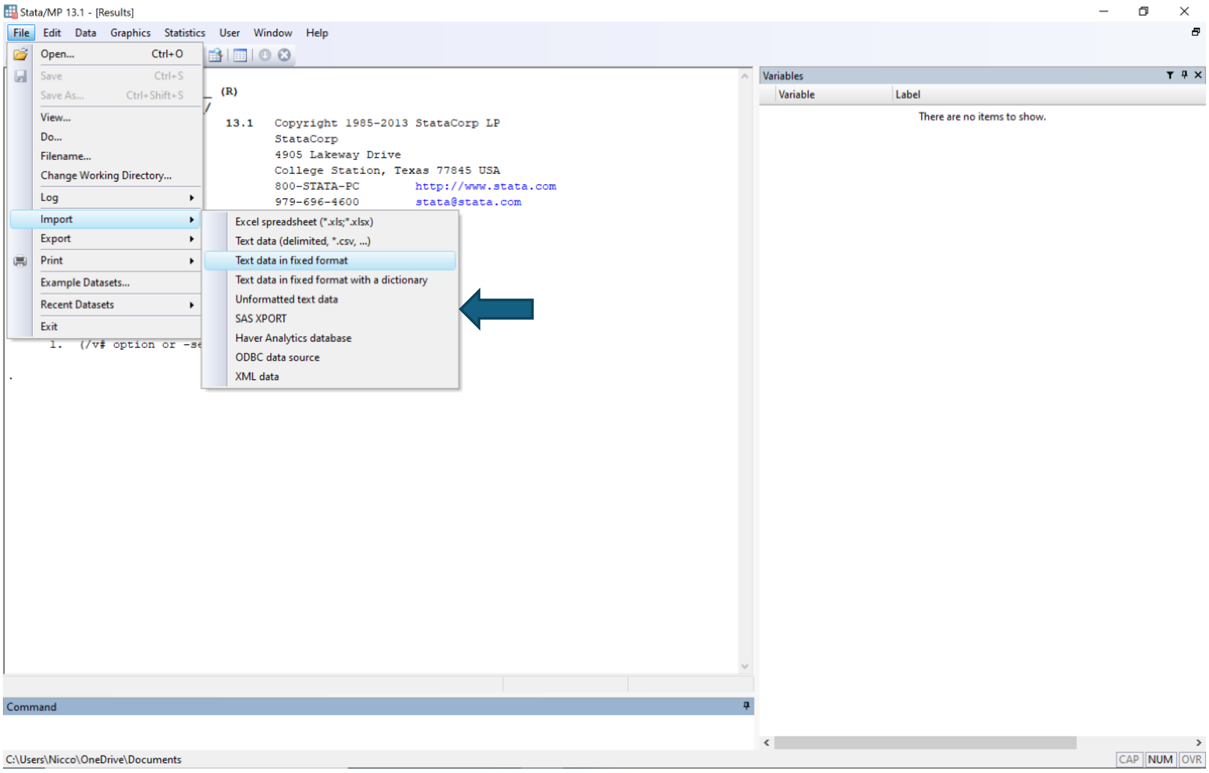
Opening an existing data set:

1: Go to "File"-> select "Open"->locate Stata file.



```
                                              (R)
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   13.1   Copyright 1985-2013 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            Colleg
                                      800-ST
                                      979-69
                                      979-69

Single-user 8-core Stata perpetual license:
       Serial number:  501306249132
         Licensed to:  Steve
                       University of Edinbur

Notes:
      1.  (/v# option or -set maxvar-) 5000 

.
```

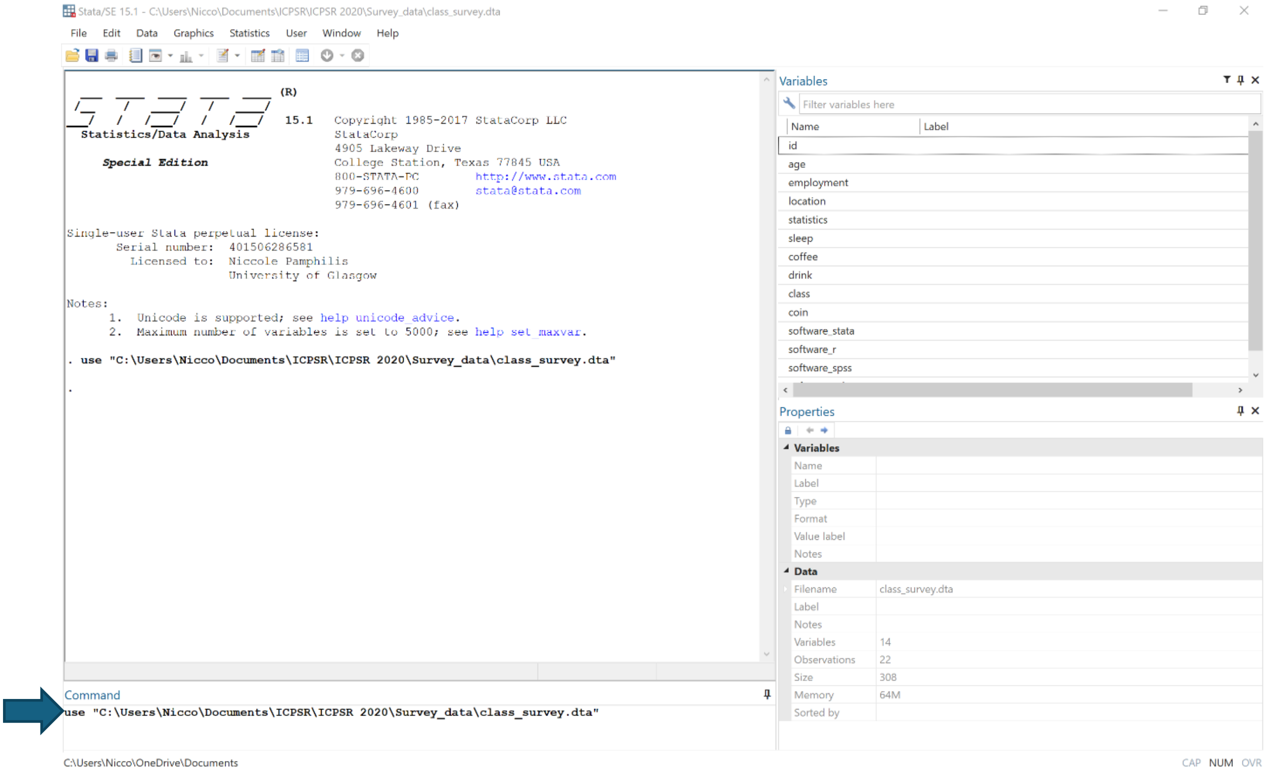
2: Go to “File”->select “Import”->select type of non-Stata file->locate file.



```
 (R)

13.1   Copyright 1985-2013 StataCorp LP
       StataCorp
       4905 Lakeway Drive
       College Station, Texas 77845 USA
       800-STATA-PC        http://www.stata.com
       979-696-4600        stata@stata.com

  1.  (/v# option or -se

.
```

Stata/MP 13.1 - [Results]
File Edit Data Graphics Statistics User Window Help
Open... Ctrl+O
Save Ctrl+S
Save As... Ctrl+Shift+S
View...
Do...
Filename...
Change Working Directory...
Log
Import
Export
Print
Example Datasets...
Recent Datasets
Exit
Excel spreadsheet (*.xls;*.xlsx)
Text data (delimited, *.csv, ...)
Text data in fixed format
Text data in fixed format with a dictionary
Unformatted text data
SAS XPORT
Haver Analytics database
ODBC data source
XML data
Variables
Variable
Label
There are no items to show.
Command
C:\Users\Nicco\OneDrive\Documents
CAP
NUM
OVR

3: Enter file pathway directly into command window



```
                                                      (R)
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  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:  401506286581
         Licensed to:  Niccole Pamphilis
                       University of Glasgow

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "C:\Users\Nicco\Documents\ICPSR\ICPSR 2020\Survey_data\class_survey.dta"

.
```

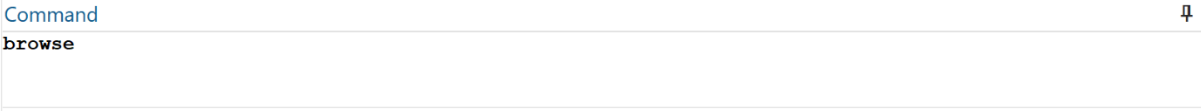
Filename class_survey.dta
Variables 14
Observations 22
Size 308
Memory 64M

```
use "C:\Users\Nicco\Documents\ICPSR\ICPSR 2020\Survey_data\class_survey.dta"
```

To view open data set:

1. Enter “edit” in your command window. Edit, allows you to make changes to the data set that you have open. Be careful when using this command because you may make changes by accident.

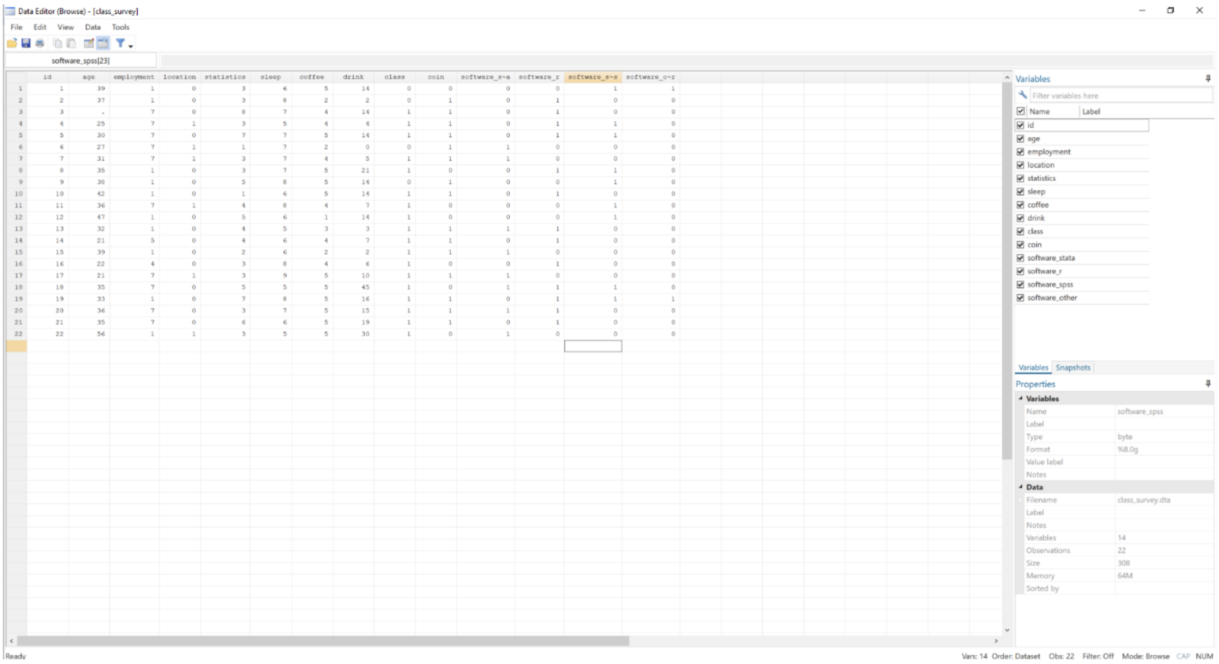
2. Enter “browse” in your command window. Browse allows you to see the data set, like the edit command, but does not allow you to make changes to the open data set.



```
browse
```

Dataset View:

Once you have used edit or browse your data set window will open in a separate window. Like the main page you have a properties window where you can edit information for the variables.



	id	age	employment	location	statistics	sleep	coffee	drink	class	coin	software_s~a	software_r	software_s~s	software_o~r
1	1	39	1	0	3	6	5	14	0	0	0	0	1	1
2	2	37	1	0	3	8	2	2	0	1	0	1	0	0
3	3	.	7	0	8	7	4	14	1	1	0	1	0	0
4	4	25	7	1	3	5	4	4	1	1	0	1	1	0
5	5	30	7	0	7	7	5	14	1	1	0	1	1	0
6	6	27	7	1	1	7	2	0	0	1	1	0	0	0
7	7	31	7	1	3	7	4	5	1	1	1	0	0	0
8	8	35	1	0	3	7	5	21	1	0	0	1	1	0
9	9	38	1	0	5	8	5	14	0	1	0	0	1	0
10	10	42	1	0	1	6	5	14	1	1	0	1	0	0
11	11	36	7	1	4	8	4	7	1	0	0	0	1	0
12	12	47	1	0	5	6	1	14	1	0	0	0	1	0
13	13	32	1	0	4	5	3	3	1	1	1	1	0	0
14	14	21	5	0	4	6	4	7	1	1	0	1	0	0
15	15	39	1	0	2	6	2	2	1	1	1	0	0	0
16	16	22	4	0	3	8	4	6	1	0	0	1	0	0
17	17	21	7	1	3	9	5	10	1	1	1	0	0	0
18	18	35	7	0	5	5	5	45	1	0	1	1	1	0
19	19	33	1	0	7	8	5	16	1	1	0	1	1	1
20	20	36	7	0	3	7	5	15	1	1	1	1	0	0
21	21	35	7	0	6	6	5	19	1	1	0	1	0	0
22	22	56	1	1	3	5	5	30	1	0	1	0	0	0

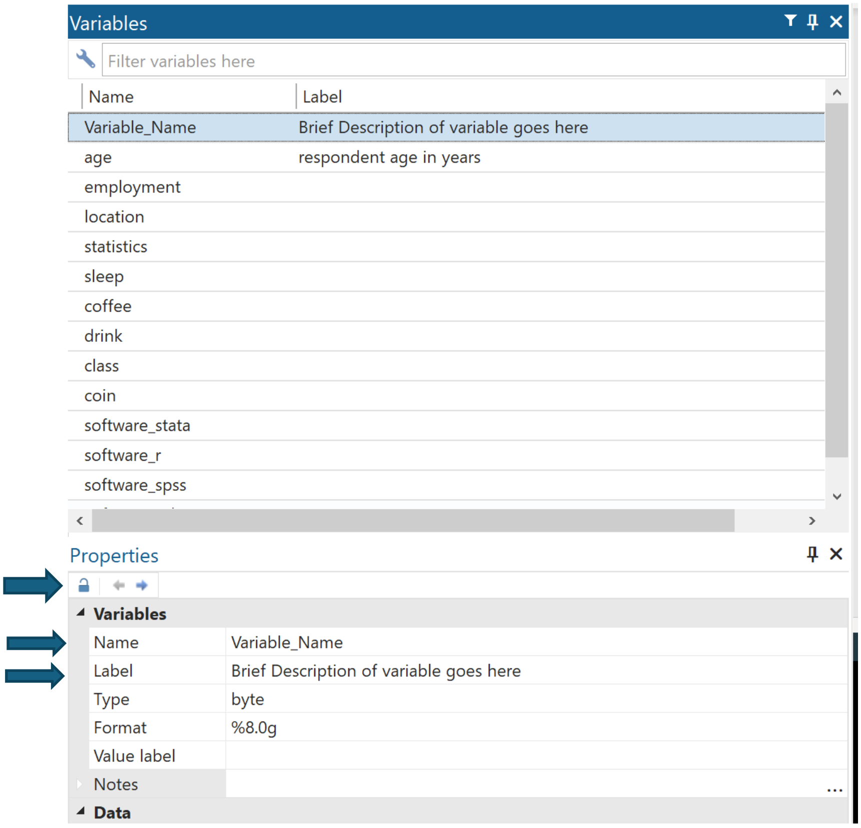
☑ Name	Label
☑ id	
☑ age	
☑ employment	
☑ location	
☑ statistics	
☑ sleep	
☑ coffee	
☑ drink	
☑ class	
☑ coin	
☑ software_stata	
☑ software_r	
☑ software_spss	
☑ software_other	

Variables	
Name	software_spss
Label	
Type	byte
Format	%8.0g
Value label	
Notes	
Data	
Filename	class_survey.dta
Label	
Notes	
Variables	14
Observations	22
Size	308
Memory	64M
Sorted by	

Vars: 14 Order: Dataset Obs: 22 Filter: Off Mode: Browse CAP NUM

Variable Details:

In the properties window (shown earlier) you can edit or add information about the variables in the data set. You can add/change variables names in the "Name" box. You can add brief descriptions to remind yourself what variables are called in the "label box. Make sure you click on the lock image so you can make changes.



Name	Label
Variable_Name	Brief Description of variable goes here
age	respondent age in years
employment	
location	
statistics	
sleep	
coffee	
drink	
class	
coin	
software_stata	
software_r	
software_spss	

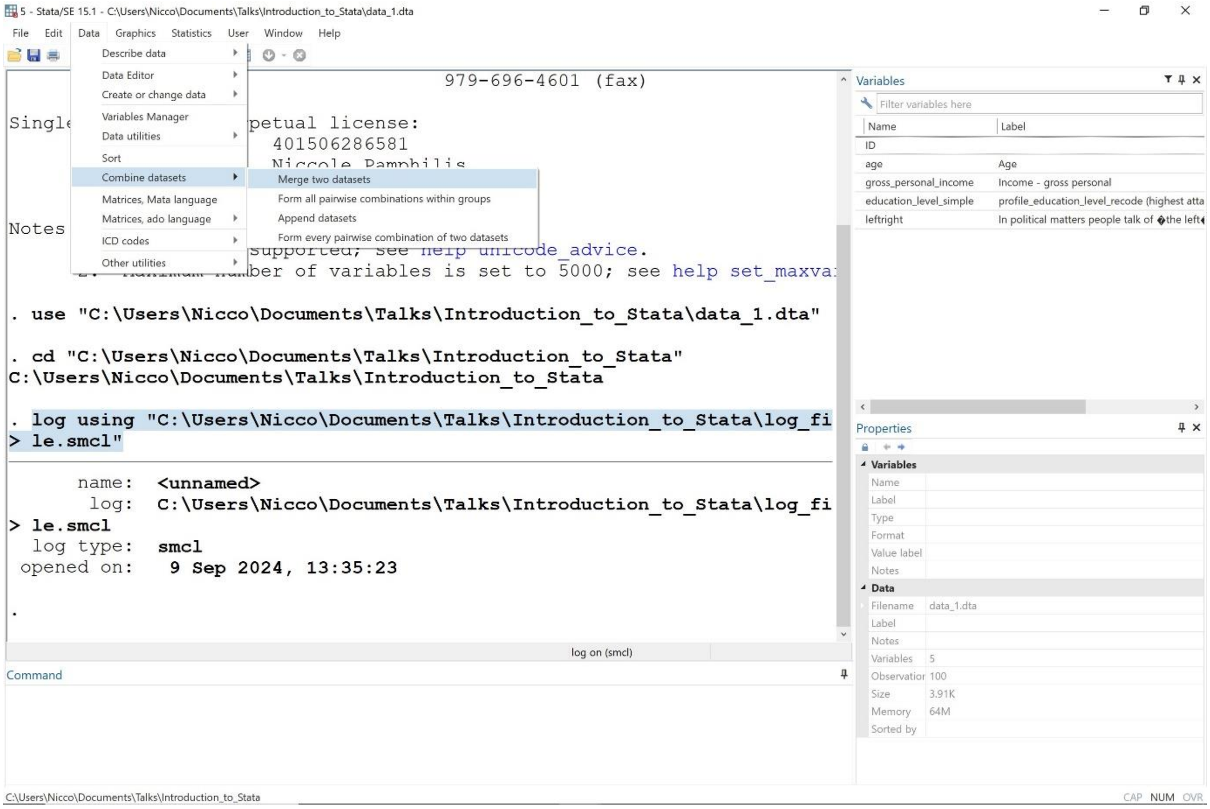
Variables	
Name	Variable_Name
Label	Brief Description of variable goes here
Type	byte
Format	%8.0g
Value label	
Notes	...
Data	

Merging Data Files

Sometimes we find new data to add to our existing data set, to account for this there are different ways to merge datasets together.

Merging Based on New Variables:

There are times when new data on existing observations is found, for example a second wave of survey data or new country-level indicators. To account for this we can merge the datasets together IF the same unique identifier exists for the observations across the two datasets. In this case we can tell Stata what the unique identified variable is and it will match and joint the datasets accordingly.



```
                                  979-696-4601 (fax)

Single      petual license:
              401506286581
              Niccole Pamphilis

Notes
            supported; see help unicode_advice.
                 ber of variables is set to 5000; see help set_maxva

. use "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\data_1.dta"

. cd "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata"
C:\Users\Nicco\Documents\Talks\Introduction_to_Stata

. log using "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\log_fi
> le.smcl"
----------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\log_fi
> le.smcl
  log type:  smcl
 opened on:   9 Sep 2024, 13:35:23

.
```

Name	Label
ID	
age	Age
gross_personal_income	Income - gross personal
education_level_simple	profile_education_level_recode (highest atta
leftright	In political matters people talk of �the left�

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Filename	data_1.dta
Label	
Notes	
Variables	5
Observatior	100
Size	3.91K
Memory	64M
Sorted by	

Append Based on New Observations:

There are other times when the data you have can be expanded to include additional observations. In this situation you would need to have the same variables across the two sets of observations upon which they are to be matched. Consider a situation where a survey is run in one city and the data is compiled, then a week later the same survey is run in another city. Same variables different observations. Now in the data tab we select append instead of merge, and tell Stata which data we want to combine it with.

append using "C:\Users\Nicco\Documents\Talks\Introduction_to_Stata\data_3.dta

Entering Data

To enter new data in Stata, open the data window using the "edit" command. Once there you can enter individual values one at a time or you can copy and paste from other files such as excel where you can highlight the column you want to add.

You can also create and edit variables using the command window

Creating a new variable and label it

*Create a new variable, called "varname"

```
Gen varname=.
```

Gen tells Stata to generate a variable.
Varname is a generic placeholder and you would type whatever you want to call the variable there.
The = tells Stata what to put in this new variable.
And the . says leave a new data points for this variable empty right now.

*Label a variable

```
label var varname "label for varname"
```

Label var tells Stata you want to create a descriptive label for the variable.
Varname is the name of the variable you want to label.
Inside the "" goes the description for the variable.

*Define value label

```
label define lblname1 "label1" 2 "label2"
```

This command allows you to associate words with the numbers (like yes or no for survey responses).
Label tells stata you are creating a description.
Define says you create a descriptors.
Lblname is what you are going to call these set of number descriptors (in case You want to use them for a few different variables with similar coding.
The numbers indicate the data values.
The text following the number in "" is the description associated with each Preceeding number.

Assign value label to a variable

```
label value varname lblname
```

The above command allows you to apply number descriptors (created in the text just above) to a variable (here one named varname).

Descriptive Statistics

You can begin by looking at the dataset that you have loaded. This is useful if you want to check how Stata is reading your variables.

Describe

```
Contains data from
C:\Users\Nicco\Documents\ICPSR\ICPSR_2021\Survey_data\class_survey.dta
  obs:            27
 vars:            14                          25 Jun 2021 10:20
 size:           459
-----------------------------------------------------------------------------------
------------------------
              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------------
------------------------
id              byte    %8.0g                 ID
age             byte    %8.0g                 Age
employment      byte    %8.0g                 Employment
location        byte    %8.0g                 Location
statistics      byte    %8.0g                 Statistics
sleep           float   %8.0g                 Sleep
coffee          byte    %8.0g                 Coffee
drink           byte    %8.0g                 Drink
class           byte    %8.0g                 Class
coin            byte    %8.0g                 Coin
software_stata  byte    %8.0g                 Software_Stata
software_r      byte    %8.0g                 Software_R
software_spss   byte    %8.0g                 Software_SPSS
software_other  byte    %8.0g                 Software_Other
-------------------------------------------------------------------
```

Question: Do you notice any issues with the type of data for Coin?

Basic Summary statistics

The sum command provides you with basic information, but assumes interval/ratio level measurement

Notice the results provide you with the number of observations, the mean, standard deviation, minimum and maximum values

To use the command start with sum then list the variables you want information on

`sum Age Sleep Location`

```
  Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         Age |         25       34.76     8.86604         22         53
       Sleep |         25        6.56    1.157584          4          8
    Location |         25         .56    .5066228          0          1
```

With the sum command you can also add the option detail, which provides you with a more complete summary of information in relation to the variable

To use the command you add, detail at the end of your command line. Notice ow that you will see the values that occur at different percentiles and quartiles, as well as information on variance and skewness.

`sum Age, detail`

```
                            Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           22             22
 5%           22             22
10%           24             24       Obs                  25
25%           30             24       Sum of Wgt.          25

50%           34                      Mean              34.76
                        Largest       Std. Dev.       8.86604
75%           38             46
90%           48             48       Variance       78.60667
95%           48             48       Skewness        .330763
99%           53             53       Kurtosis        2.18622
```

The tab command provides you with a frequency table for the variable

Notice the results provide you with the count for each category where there was at least one observation, a relative frequency and a cumulative frequency. Using this you can locate the median and mode.

To use the command, start with tab and then select your variable

```
tab Coffee
```

```
     Coffee |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        8.00        8.00
          2 |          1        4.00       12.00
          3 |          4       16.00       28.00
          4 |          5       20.00       48.00
          5 |         13       52.00      100.00
------------+-----------------------------------
      Total |         25      100.00
```

Question: Do we observe all the possible values for the variable?

What is our best guess for the value an observation takes on for “Coffee”? How did you determine this?

Using Stata as a Calculator

It is possible to use Stata as a calculator as well using the di command.

To use the command start the line with di then use +-*/^ as well as keeping in mind order of operations

```
      di 7+3

             10
```

Basic Graphics/ Data Visualisations

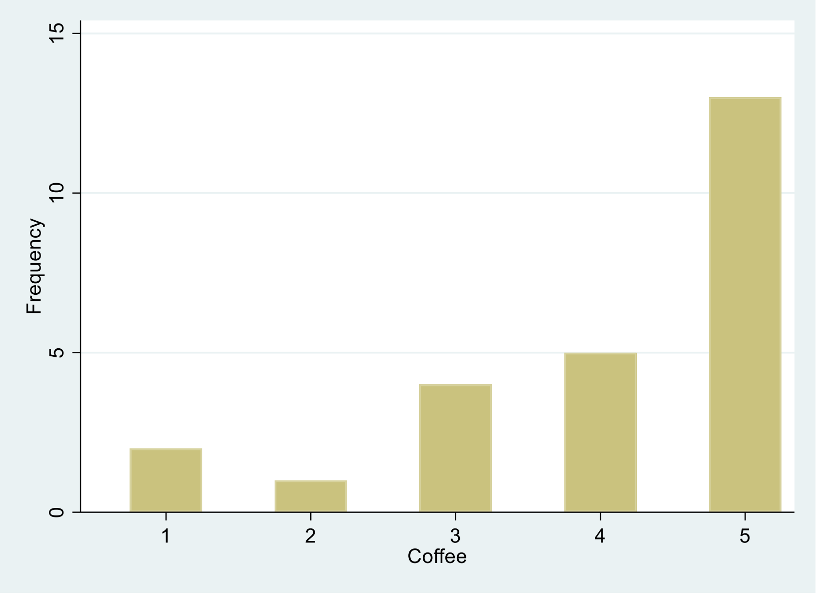
Bar charts

When you are working with ordinal or nominal level data and you want to see how values are distributed across the categories you can use a bar chart

The command is hist with the option discrete applied. To use the command type hist then the variable then, discrete

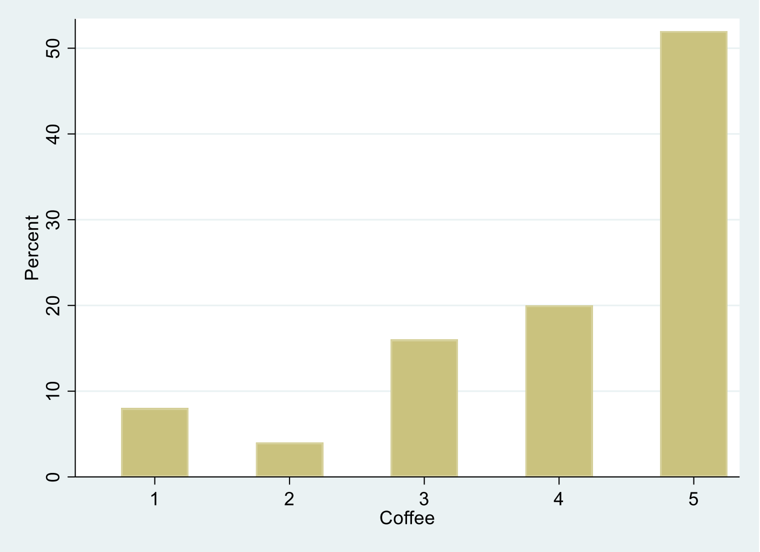
```
hist Coffee, frequency discrete width(.5)
```

```
        *The frequency option gives you a count on the y-axis
```



```
hist Coffee, percent discrete width(.5)
```

```
*The percent option gives you the proportion of
observations on the y-axis
```

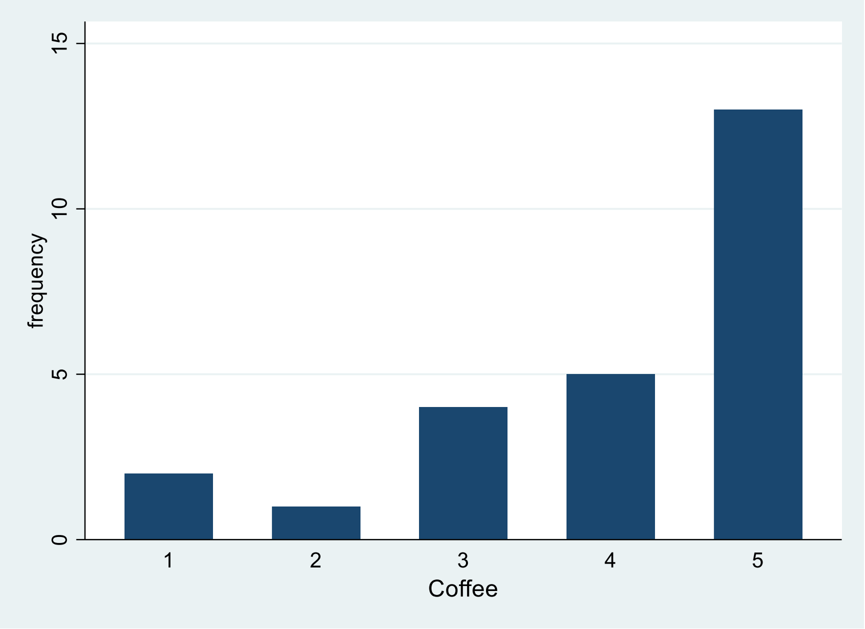


Another option is to install the catplot function (only need to do this one time)

```
ssc install catplot
```

Then you do not need to put in the gaps between the bars, note I use the recast(bar) option to I have vertical not horizontal bars

```
catplot Coffee, recast(bar)
```

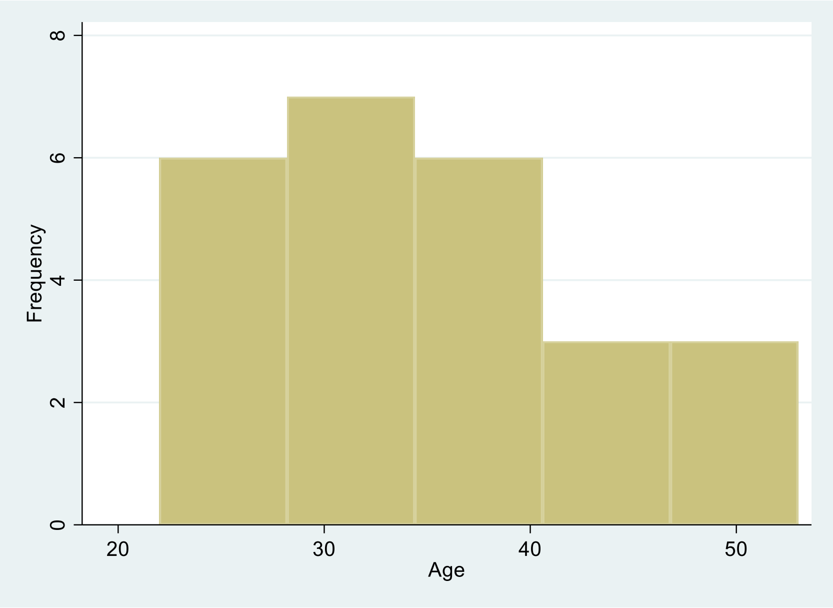


Histograms

Similar to the command for creating a bar chart to create a histogram in Stata we use the hist command.

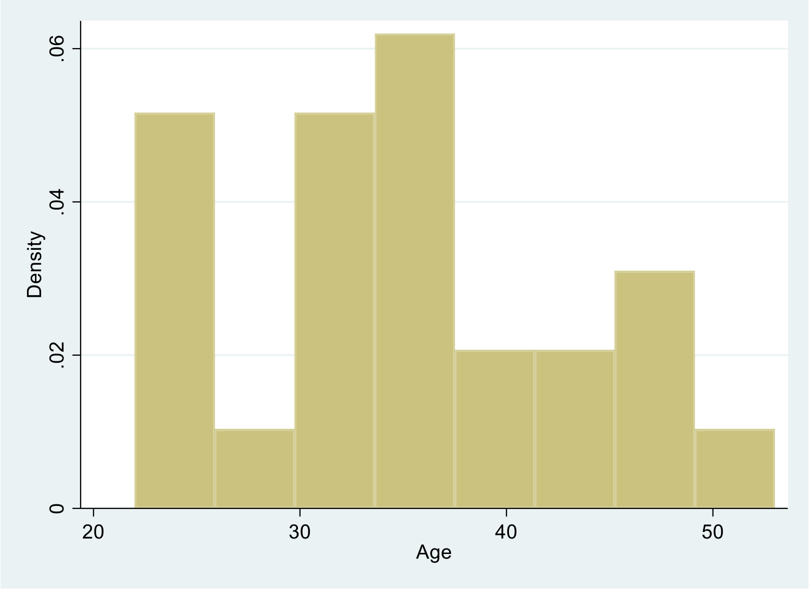
As with earlier you use the hist command first followed by your variable

```
hist Age, freq
```

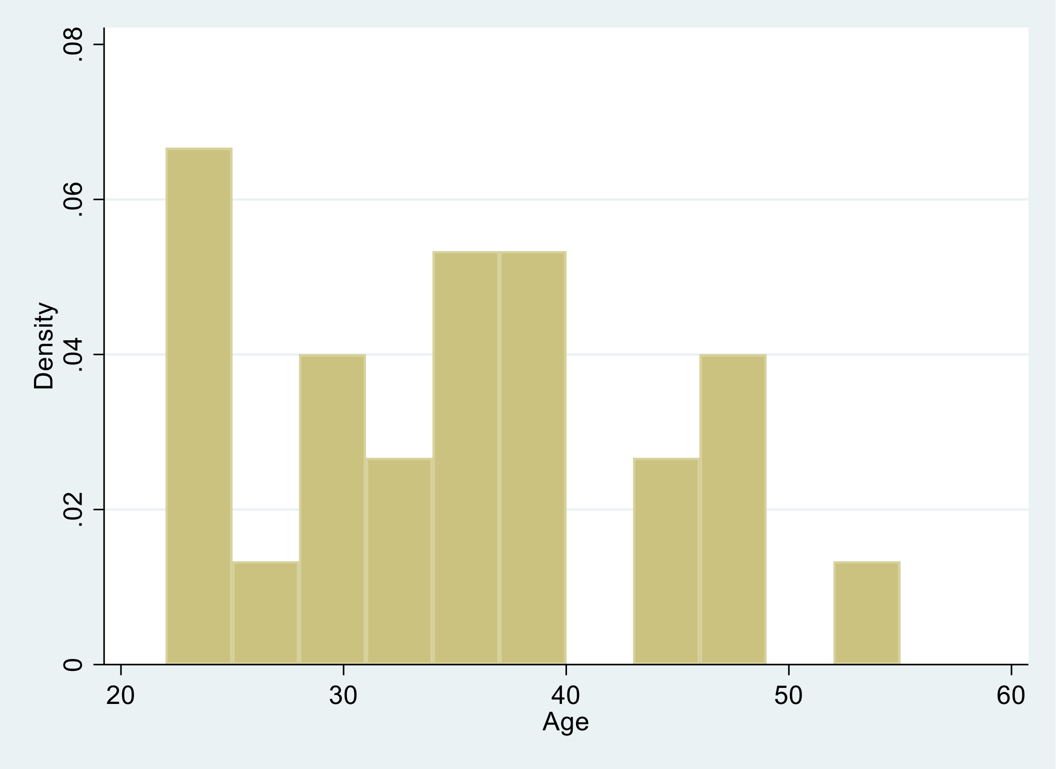


You can also play around with the number of bins(number of bars in the graph) or bin size (number of values in each bin) to adjust your graph.

```
hist Age, bin(8)
```



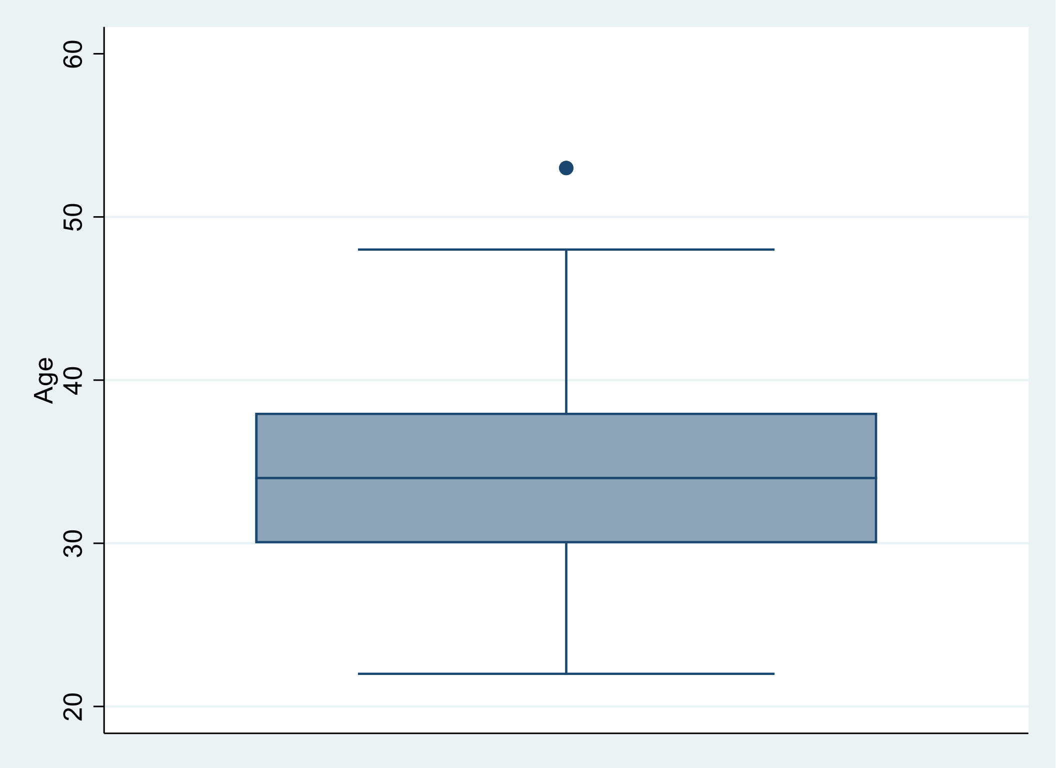
```
hist Age, width(3)
```



Boxplots

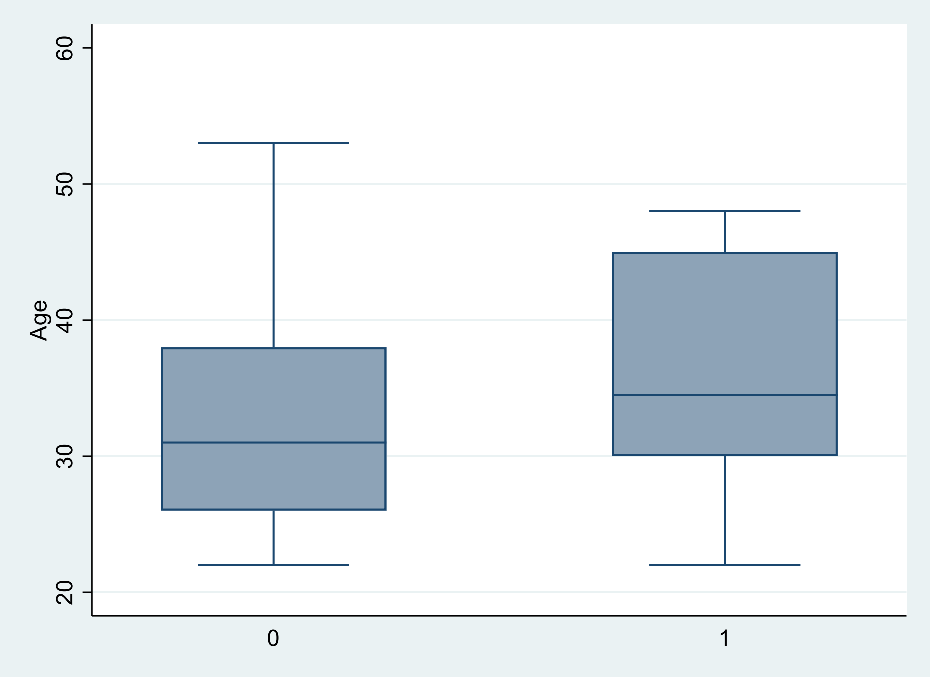
To generate a basic boxplot you can use the command graph box followed by the variable

```
graph box Age
```



You can also examine distributions across other variables with the over()option

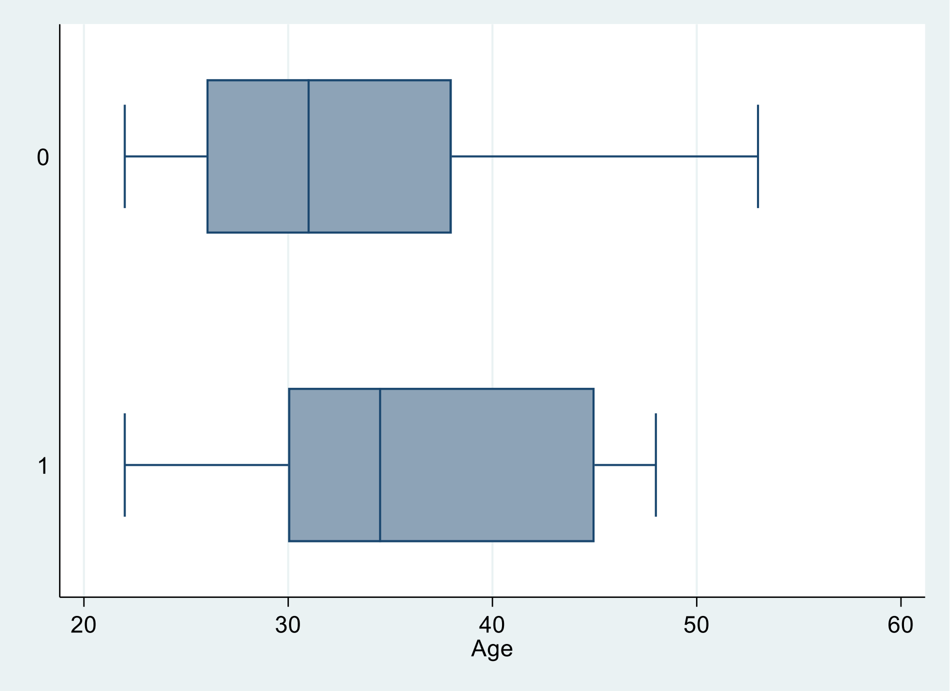
```
graph box Age, over(Location)
```



Notice that the basic coding does not apply labels to 0/1 which are needed to read it.

You can also rotate the box from being vertical to horizontal

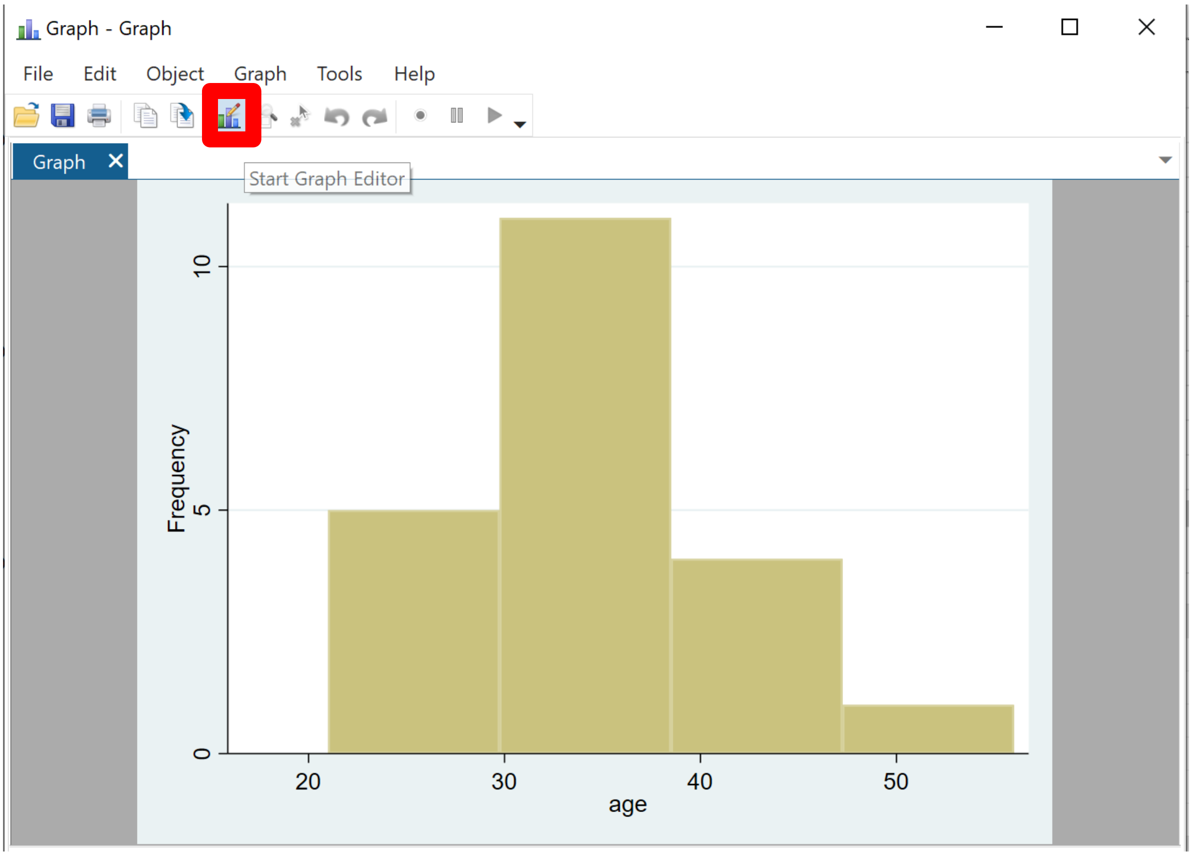
```
graph hbox Age, over(Location)
```



Changing the Look of your Graph

You can edit the appearance of your graph using two different approaches, you can use the graph editor or you can use command when creating your graph.

1. Graph editor can be found on the image of the graph once it has been created. Keep in mind you will need to save the changes and if you recreate the graph you will need to recreate these changes by hand again.



2. You can use commands in your line of code for your graph. Examples include

   To add a y-axis title

   ```
   ytitle("Systolic blood pressure")
   ```

   To add an x-axis title

   ```
   xtitle(“”)
   ```

   To add a main title to the graph

   ```
   title("")
   ```

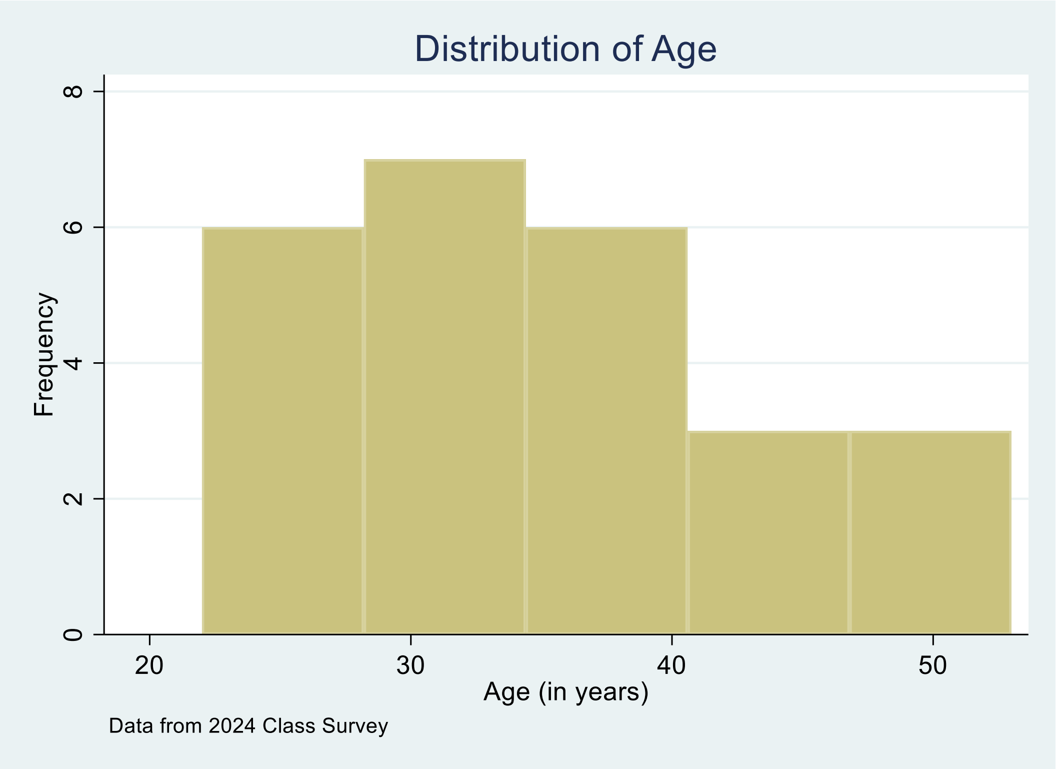
   To add a subtitle

   ```
   subtitle("()" " ")
   ```

   To add a note at the bottom

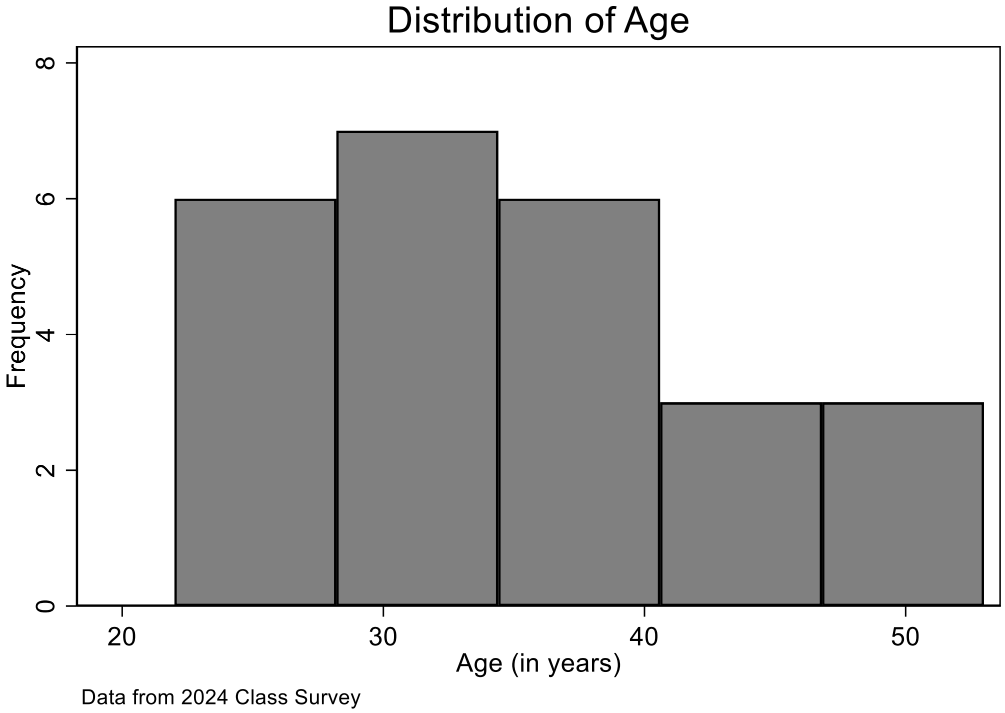
   ```
   note("")
   ```

```
    hist Age, freq xtitle("Age (in years)")
ytitle("Frequency") title("Distribution of Age") note("Data
from 2024 Class Survey")
```



You can add pre-set color schemes using the set scheme command (use help set scheme to see options)

```
     set scheme s1mono
     hist Age, freq xtitle("Age (in years)")
ytitle("Frequency") title("Distribution of Age") note("Data
from 2024 Class Survey")
```



Basic Univariate Test commands/ Command Structure

Some commands in state are used for exploring univariate statistics, or characteristics in relation to a single variable, similar to the descriptive statistics commands we just explored.

A few more examples are demonstrated below.

One-Sample T-Test

A one-sample t-test compares the mean for a sample variable to some expectation you might have about the mean in the population. Stata runs the non-directional and both directional tests at once.

```
Note a one-sample t-test requires a continuous variable (i.e., interval or
      ratio level data)
**The basic command is ttest followed by the variable you are looking at.
      You will then set the variable equal to whatever value is in your null
      hypothesis
**Note, that we use a double = not just one
```

```
ttest Age==18
```

```
One-sample t test
------------------------------------------------------------------------------
```

```
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |      21    34.14286    1.854577    8.498739    30.27428    38.01144
------------------------------------------------------------------------------
    mean = mean(age)                                              t =   8.7043
Ho: mean = 18                                    degrees of freedom =       20

    Ha: mean < 18               Ha: mean != 18                Ha: mean > 18
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Two-Sample T-Test

A two-sample t-tests compare the mean for a sample variable across two sub-groups (instead of all observations at once like the one-sample t-test). As with the one-sample t-tests the two-sample t-test runs the non-directional and directional test at the same time. Just be aware of which group it treats as group 1 and which as group 2.

By default, Stata assumes your two groups have equal variance. Therefore, if you do not tell it otherwise, this is how it will calculate the standard errors.

```
**Note that a two-sample t-test uses the same basic ttest command as above
**Now you will include an options with your grouping variable (this variable
      needs to have only two options, otherwise you cannot run a two-sample
      t-test and instead need to use an ANOVA

ttest Age, by(Location)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      15    34.73333    1.747016    6.766162    30.98636    38.48031
       1 |       6    32.66667    5.116422    12.53262    19.51448    45.81885
---------+--------------------------------------------------------------------
combined |      21    34.14286    1.854577    8.498739    30.27428    38.01144
---------+--------------------------------------------------------------------
    diff |            2.066667    4.185159               -6.692972    10.82631
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.4938
Ho: diff = 0                                     degrees of freedom =       19

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6864         Pr(|T| > |t|) = 0.6271          Pr(T > t) = 0.3136
```

If you want Stata to assume unequal variance, then you will need to use the option “unequal” as shown below.

```
**If you want to run a two-sample t-test assuming unequal variance you will
      need to tell Stata with the option "unequal"

ttest Age, by(Location) unequal
```

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      15    34.73333    1.747016    6.766162    30.98636    38.48031
       1 |       6    32.66667    5.116422    12.53262    19.51448    45.81885
---------+--------------------------------------------------------------------
combined |      21    34.14286    1.854577    8.498739    30.27428    38.01144
```

```
---------+--------------------------------------------------------------------
    diff |            2.066667    5.406463              -11.05786     15.1912
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.3823
Ho: diff = 0                     Satterthwaite's degrees of freedom =  6.20375

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6425         Pr(|T| > |t|) = 0.7150          Pr(T > t) = 0.3575
```

The above code format works when you have a single variable that captures the mean and a single variable that captures the groups. In the situation above the age variable captured age for everyone, while location indicated which group a respondent belonged to. However, sometimes our data are broken up differently, consider the variable life expectancy. We could have country level data were there is a variable for male life expectancy by country and one for female life expectancy by country. In this situation, the code above would not work and instead we would need to use the format below.

```
ttest varname1 == varname2
```

In the code above, varname1 would represent male_life_exp and varname2 would represent female_life_exp. For other more unique situations check out the help ttest file.

F-test

The ratio of variance test in Stata allows you to compare the variances for two groups and can be done using the code below. Remember, you need to do this before you can carry out a two-sample t-test.

```
**To run a ratio of variance test we use the sdtest command.
**To run the test you need to know which variable represents your continuous
      variable and which represents you group variable
**The continuous variable is used first and the grouping variable is used in
      the option "by" command

sdtest Age, by(Location)
```

```
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      15    34.73333    1.747016    6.766162    30.98636    38.48031
       1 |       6    32.66667    5.116422    12.53262    19.51448    45.81885
---------+--------------------------------------------------------------------
combined |      21    34.14286    1.854577    8.498739    30.27428    38.01144
------------------------------------------------------------------------------
    ratio = sd(0) / sd(1)                                         f =   0.2915
Ho: ratio = 1                                    degrees of freedom =    14, 5

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0312         2*Pr(F < f) = 0.0624           Pr(F > f) = 0.9688
```

Optional Commands Changes

The default for most commands include a basic set-up with assumptions about what you are seeking to do, however, most commands come with additional optional commands that allow you to either alter the default assumptions run by the basic code or to include additional steps within a single line of code. Optional commands typically follow the basic setup after a “,”.

A few examples are included below.

Confidence level

Stata assumes a 95% confidence level when calculating confidence intervals. If you would prefer a 90% or 99% confidence level you can change it with the optional command below.

```
       **By default Stata provides 95% confidence intervals for its estimates.
       **To change this you can use the "level" option after many commands
```

```
**95%
       ttest Statistics==1
```

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
statis~s |      22    3.909091    .3942723    1.849301    3.089157    4.729025
------------------------------------------------------------------------------
    mean = mean(statistics)                                       t =   7.3784
Ho: mean = 1                                     degrees of freedom =       21

    Ha: mean < 1                 Ha: mean != 1                 Ha: mean > 1
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

```
**90%
       ttest Statistics==1, level(90)
```

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
statis~s |      22    3.909091    .3942723    1.849301     3.23065    4.587532
------------------------------------------------------------------------------
    mean = mean(statistics)                                       t =   7.3784
Ho: mean = 1                                     degrees of freedom =       21

    Ha: mean < 1                 Ha: mean != 1                 Ha: mean > 1
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

```
**99%
       ttest Statistics==1, level(99)
```

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
statis~s |      22    3.909091    .3942723    1.849301    2.792764    5.025417
------------------------------------------------------------------------------
    mean = mean(statistics)                                       t =   7.3784
Ho: mean = 1                                     degrees of freedom =       21

    Ha: mean < 1                 Ha: mean != 1                 Ha: mean > 1
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Bivariate Commands

Bivariate commands move our exploration of data further and allow for statistical analysis of pair-wise relationships across variables.

Cross-Tabulation

To generate a cross-tab you need to select two variables that have discrete categories (think: nominal and ordinal level data). The resulting table will show the pair-wise frequencies for observations across these two variables.

```
       *Basic command is tab (same as with frequency table)
       *Next select the two variables for comparison

         tab Location Class
```

```
           |         Class
  Location |         0          1 |     Total
-----------+----------------------+----------
         0 |         0         11 |        11
         1 |         2         12 |        14
-----------+----------------------+----------
     Total |         2         23 |        25
```

Commands within Commands, Chi-2 Test

To test for the existence of a relationship between two variables in a cross-tab you can use a chi-2 test that looks at how the observed pair-wise frequencies deviate from what would be expected if the two variables were not related to one another.

```
       **To run the chi2 test you need to run the test with the cross-tab
       **You will use the option "chi2"

         tab Location Class, chi2
```

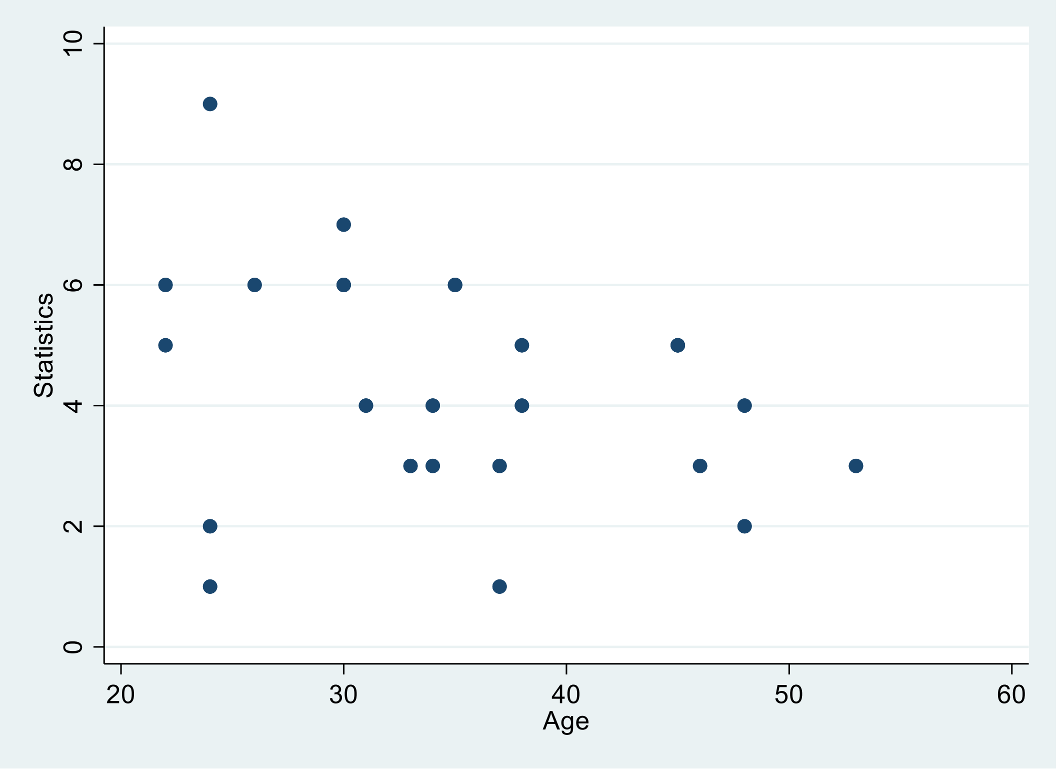
```
           |         Class
  Location |         0          1 |     Total
-----------+----------------------+----------
         0 |         0         11 |        11
         1 |         2         12 |        14
-----------+----------------------+----------
     Total |         2         23 |        25

          Pearson chi2(1) =   1.7081   Pr = 0.191
```

Scatter Plot

To visually assess the relationship between two continuous level variables (think: interval and ratio level variables) you can use a scatter plot. The graph with plot the pair-wise values for each observation.

```
**scatter plots use the graph command
**within graph you are creating a "twoway graph" with two variables
**The type of graph is scatter
**Then you list your dependent variable followed by the independent variable
```

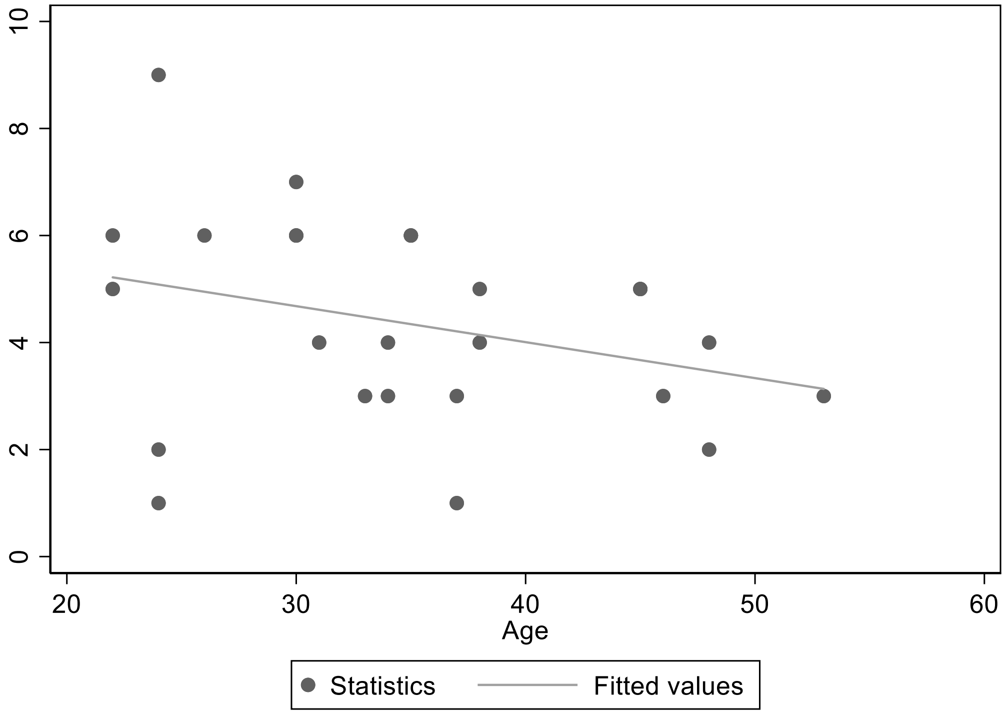
```
graph twoway (scatter statistics age)
```



You can also add a linear line to see how the points follow a straight line.

```
**Scatter with linear line
        **To add a linear best fit line, you add another graph over it
```

```
graph twoway (scatter statistics age) (lfit statistics
age)
```

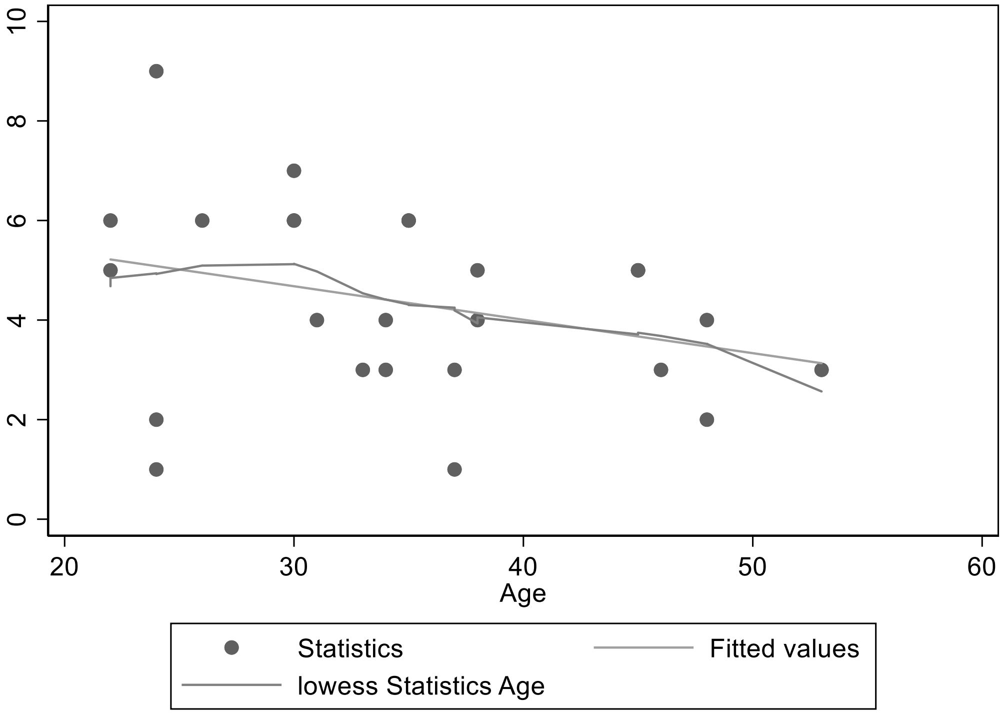


You can also add a loess curve to see what the actual pattern of the observations follows.

**Scatter with Loess curve

```
**Scatter with Loess curve
```

```
     graph twoway (scatter statistics age) (lfit statistics
age)(lowess statistics age)
```



Correlation

To test the strength of the linear relationship between the two variables you can calculate the correlation value. The value will range between -1 and 1, which sign and value indicating direction and strength of the relationship.

```
      **to generate a basic correlation value you can use the corr command
      **next you select your two (or more variables)
```

```
          corr Statistics Age
(obs=25)

             | Statis~s      Age
-------------+------------------
  Statistics |   1.0000
         Age |  -0.3082   1.0000
```

```
          corr Statistics Age Drink
(obs=25)

             | Statis~s      Age    Drink
-------------+---------------------------
  Statistics |   1.0000
         Age |  -0.3082   1.0000
       Drink |  -0.0855   0.1189   1.0000
```

. You can also test how this value deviates from no relationship (r=0).

```
      **You can also use the pwcorr option which is for pairwise correlations and
            this has the option for a chi2 test
      **The chi2 test uses the option "sig"
```

```
  pwcorr Statistics Age, sig

             | Statis~s      Age
-------------+------------------
  Statistics |   1.0000
             |
             |
         Age |  -0.3082   1.0000
             |   0.1339
             |             |
```

Modelling Relationships

Modelling relationships requires a clear question and understanding of the variables you want to include and the requirements of the model you want to run. For example, do you have enough observations to run the model? Or, do you have the right types of variables?

Below is the basic setup for and OLS regression model, which is similar in design to a lot of model set-ups.

Regression Model

Once you move on to modelling, you may consider running an OLS regression model. The basic set-up below is for a bivariate regression (one dependent and one independent model). But the code is the same if you have multiple independent variables.

```
       **The basic command to execute an OLS Regression model is "reg"
       **after reg the next variable listed is assumed to be the outcome/dependent
variable/Y
       **The next variable(s) represent the independent variable(s).

          reg Statistics Age Drink Sleep
```

```
      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(3, 21)        =      2.25
       Model |  21.8478283         3  7.28260943   Prob > F        =    0.1121
    Residual |  67.9121717        21  3.23391294   R-squared       =    0.2434
-------------+----------------------------------   Adj R-squared   =    0.1353
       Total |       89.76        24        3.74   Root MSE        =    1.7983

------------------------------------------------------------------------------
  Statistics |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |  -.0427463   .0432623    -0.99   0.334    -.1327153    .0472226
       Drink |  -.1119342   .0832063    -1.35   0.193     -.284971    .0611027
       Sleep |   .7817063   .3883581     2.01   0.057    -.0259286    1.589341
       _cons |   1.667071   2.980297     0.56   0.582    -4.530796    7.864939
```

Help Command

As you start to create graphs you will have situations unique to your data and the story you want to capture with a graph, which means you will need to find the right set of commands for you to create that graph. You can start by checking out the Stat graph help files:

https://www.stata.com/features/publication-quality-graphics/

If you know the specific command you want to use but are interested in seeing the options, you can type help and the command directly in the State command window

```
       help mlogit
```