

# STATA CLASS NOTES

“BEHAVIORAL ECONOMICS AND CONSUMER DECISION MAKING”

(Laurea Magistrale Analisi e Misure di Marketing)

September 26, 2023

Veronica Pizziol

[veronica.pizziol@unibo.it](mailto:veronica.pizziol@unibo.it)

## Stata basics

### Before starting

- This course is not about how to download and install Stata. You should already have installed it on your laptop with the legal license that Luiss grants you!
- No prior exposure to computer programming is required to successfully attend this class.

### Introduction

These notes and the “do-file” of today’s class will provide you with the materials necessary to have an introductory knowledge about Stata (version 18 or any other fairly recent version) for data management and statistical analysis. You can use them in combination with the facilities that Stata provides you to know more about specific commands:

- For information on subject or procedure “y” type “help y” into the Command window (or go to the help “menu” with the mouse).
- Another very useful command is “findit”. If you look on information on “y” type “findit y” into the Command window. This makes Stata search the web as well as the “help” information installed with the software.

These materials will be enough for this course. However, if you want to have more information on Stata, some useful resources are:

- The UCLA computing service website on Stata is very useful resource: <https://stats.oarc.ucla.edu/stata/>. There is a large range of advice on different aspects of Stata here, including introductory guides that can be downloaded. This is definitely a website worth looking at when you are first starting to use Stata!
- Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics Using Stata* (Vol. 2). College Station, TX: Stata press. This book is useful especially for economic analysis, but many issues, models and methodologies discussed are also relevant for other social sciences. However, this book is quite advanced, and **not needed for this course**, so just keep it in mind as a reference for the future if you plan to deepen your quantitative analytical skills.

# What is Stata?

- It is a multi-purpose data analysis software widely used in social science and economic research.
- New releases of Stata are distributed annually to offer a wide array of statistical tools including newer advanced methods.
- Stata helps you **explore, manage, summarize, and analyze datasets**.
- What is a dataset? It is a collection of pieces of information called *variables*. Variables are the key means to manipulating data. They are usually arranged by columns. Each variable can have one or several *values*.

## Stata interface

- **(A) Command window:** it is where you type commands that can be launched by pressing ENTER.
- **(B) Variables window:** it lists the variables (and their labels) in memory. Clicking on a variable name will make its description appear in the Properties Window, while double-clicking on it will make it appear in the Command Window.
- **(C) Properties window:** it shows information on the variables and dataset.
- **(D) Results window:** it displays the outputs from commands launched in the Command window (or through the script/dofile).
- **(E) Review window:** it lists the commands launched from the Command window. Successful commands will appear black, while unsuccessful ones appear red. You can launch again a command by double-clicking on the Review window.

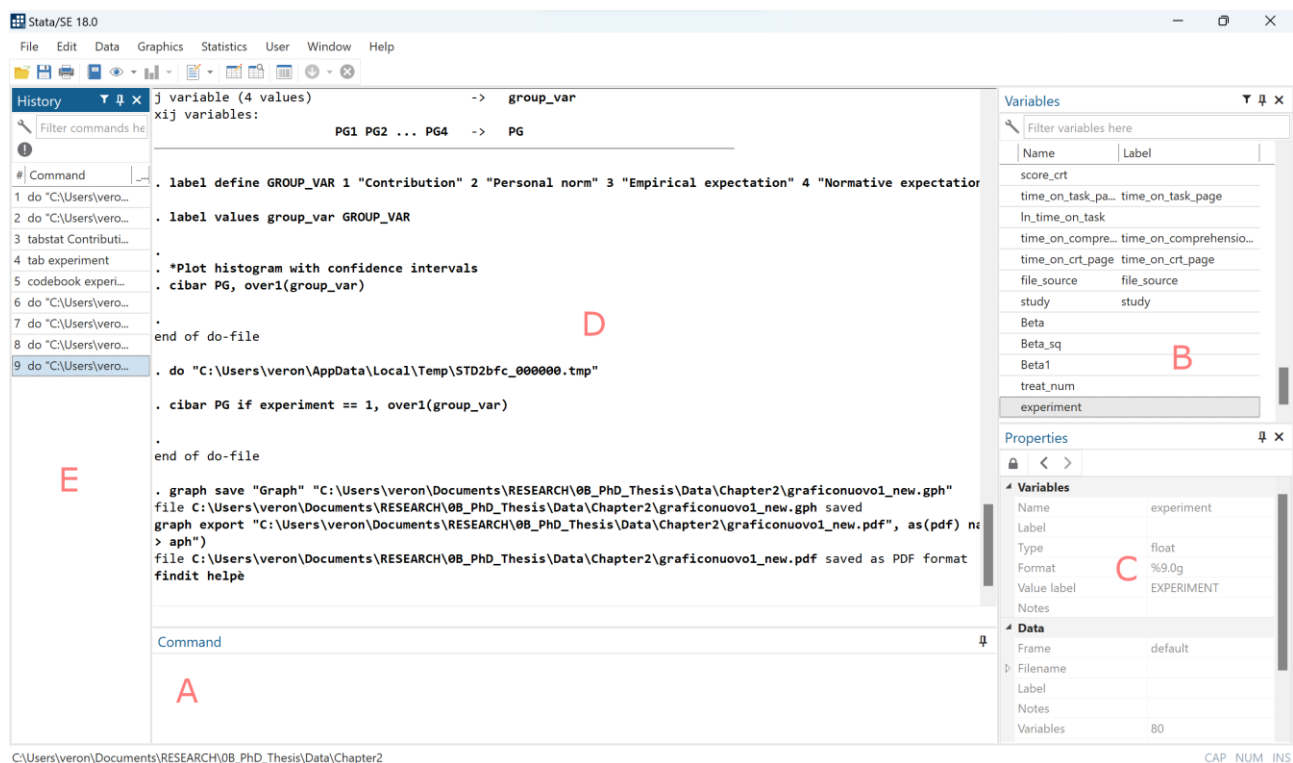


Figure 1: Stata screen.

## Stata menus

- We will mostly use *syntax* to run commands (as most Stata users do).
- Syntax refers to the set of rules that define the structure and format of commands in a programming language.
- Stata however allows for a “point-and-click” way to run commands through its menus.
- Stata menus can run most of its data management, graphical, and statistical commands.

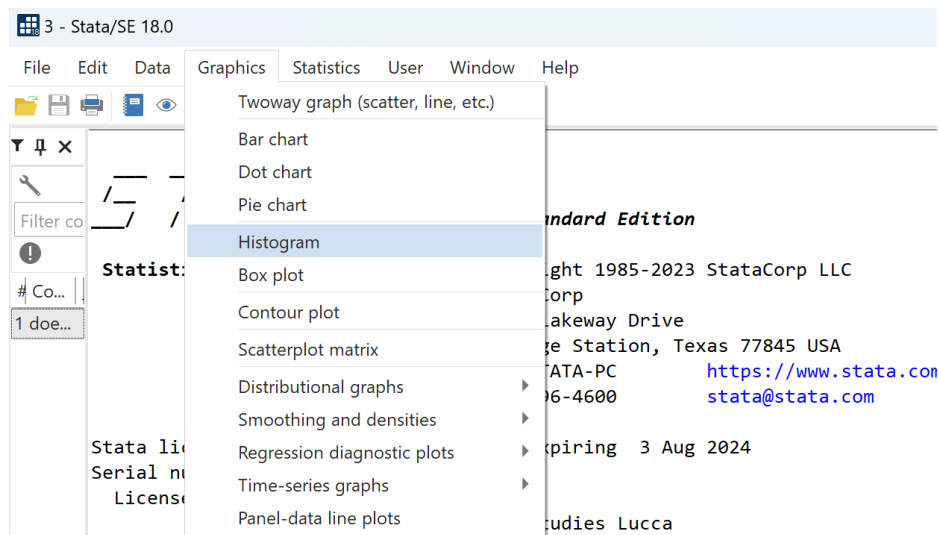


Figure 2: Creating a histogram through the Stata menu “Graphics”.

## Main Stata files

- **.dta**: it contains data that can be loaded directly in Stata without import functions.
- **.do**: do-files are the script files containing a list of commands.

« MLPGG » Project_01 » Analysis »			Cerca in Analysis
Nome	Ultima modifica	Tipo	
00_Analysis_Aug2020	16/05/2023 21:47	Cartella di file	
00_MLPGG	16/05/2023 21:47	Cartella di file	
3efficiency_tests_2	04/11/2020 18:58	Stata Do-file	
avgRC1	10/11/2020 11:05	File TEX	
Dataset_VP_1	01/09/2020 15:20	Stata Dataset	
Dataset_VP_5	02/09/2020 00:15	Stata Dataset	

Figure 3: How “.dta” file and “.do” files look in a folder.

## Do-file

- Do-files are text files where you can **save and run your commands for reuse**, rather than retyping the commands every time in the Command window.
- To open a new do-file editor, there are two ways:
  - Write the command *doedit* in the Command window
  - Or click on the “paper and pencil” icon on the toolbar (see Figure 4 below).

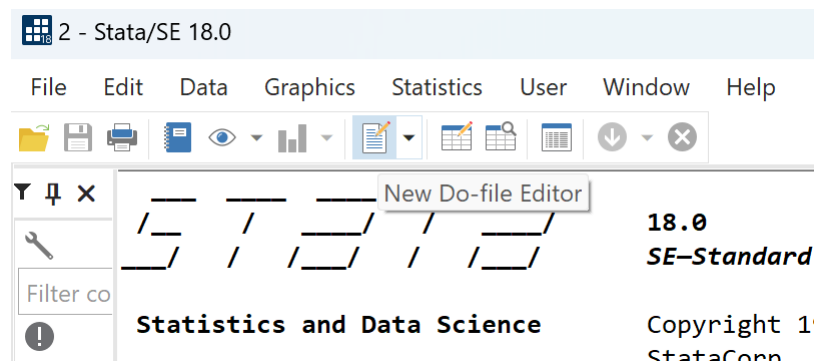


Figure 4: Open a new do-file editor.

- To execute your commands, you need to highlight the part of the command you want to run, and then hit Ctrl+D (or, if you have a Mac: Shift+Cmd+D) or to click on the “Do” icon (as in Figure 5 below). Multiple commands can be selected and executed!

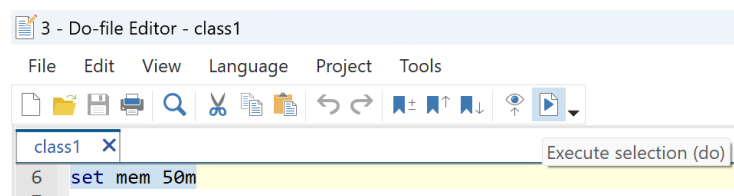


Figure 5: The “Do” icon.

- You can comment your code by using the characters `/*` before and `*/` after the text. You can also use asterisk(s) or `/**` to comment a single line. Comments will appear colored green. They are not meant to be read by the software (they are NOT commands) but by HUMANS. Adding comments to your code will help others (and the “future you”) to interpret it!

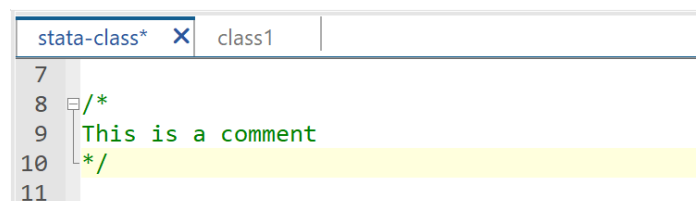


Figure 6: Example of a comment.

## Load your data

- FIRST always set the current directory where you are working from your laptop (that is, the “folder” where you have your dataset and/or where you will save it). You can do this with the `cd` command.
- Then you need to load the specific dataset. If it is in “dta” format you can use the `use` command.

```
19  **This command sets your current directory
20  cd "PUT THE DIRECTORY WHERE YOU HAVE THE DATA HERE"
21
22  **This command loads the dataset that you have in "dta" format
23  use dataset_name.dta
```

Figure 7: Set the directory and load the dataset.

- If you want to delete everything from memory (and this is ALWAYS the case when you start a new do-file), you need to specify the option `clear` after a comma, on the same line of the `use` command. Your command line will then look like this: `use dataset_name.dta, clear`
- If the dataset is NOT in “dta” format, but for instance in an Excel or csv format, you need to IMPORT the dataset (the `use` command is not useful in this case). To serve this purpose, use the `import` command. You can specify the *options* for the command after the comma (this is true for every command). Figure 8 shows an example of import with Excel file. Figure 9 shows an example of import with csv file.

```
26  **This command imports the dataset from other formats (e.g., xlsx)
27  import excel "dataset_name.xlsx", sheet("Sheet1") firstrow clear
```

Figure 8: Import Excel file (example).

```
15  ** This command imports the dataset from other formats (e.g., csv)
16  import delimited "dataset_name.csv", clear
```

Figure 9: Import csv file (example).

- If you are struggling with these commands, remember! There is the “point-and-click” way to import files with the menu available **File >> Import** (see Figure 10 below).

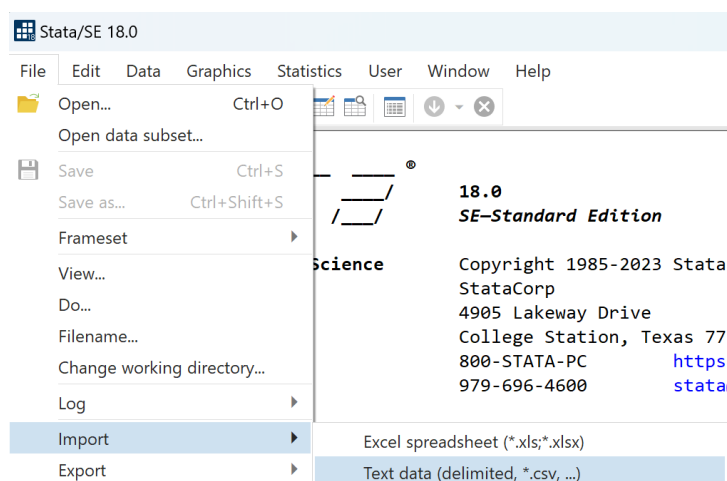


Figure 10: Import files from other formats via the “File” menu from the main Stata interface.

- To know more about how to use the “import” command, **type “help import” in the Command window.**
- In this class, we will use a “.dta” dataset that Stata can retrieve over the Internet, so you don’t need to have it stored on your PC. There are two commands that can do this. They are the *sysuse* and the *webuse* commands. We will use it just in this class, and for sure you won’t need them in the assignments.
- After using the do-file, remember to save it by clicking on the floppy disk icon. The asterisk in the file name means that there are unsaved changes!

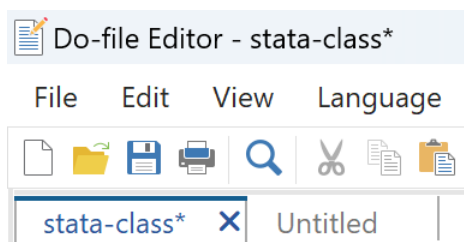


Figure 11: How to save changes on the do-file.

## Browse your data

- To open and visualize the dataset, you can:
  - Write the command *browse*
  - Or click on the “spreadsheet with magnifying glass” icon on the toolbar.

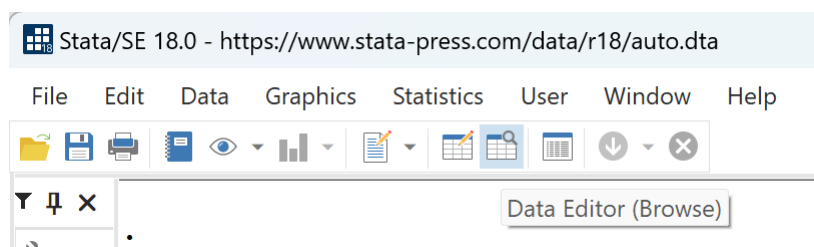



Figure 12: Browse the dataset.

- You will notice different colors for different data types. Black columns are numeric, red columns are strings, and blue columns are numeric with string labels.

foreign[60]											1	
	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displac...	gear_ratio	foreign
45	Plym. Sapporo	6,486	26 .		1.5	8	2,520	182	38	119	3.54	Domestic
46	Plym. Volare	4,060	18 Fair		5.0	16	3,330	201	44	225	3.23	Domestic
47	Pont. Catalina	5,798	18 Good		4.0	20	3,700	214	42	231	2.73	Domestic
48	Pont. Firebird	4,934	18 Poor		1.5	7	3,470	198	42	231	3.08	Domestic
49	Pont. Grand Prix	5,222	19 Average		2.0	16	3,210	201	45	231	2.93	Domestic
50	Pont. Le Mans	4,723	19 Average		3.5	17	3,200	199	40	231	2.93	Domestic
51	Pont. Phoenix	4,424	19 .		3.5	13	3,420	203	43	231	3.08	Domestic
52	Pont. Sunbird	4,172	24 Fair		2.0	7	2,690	179	41	151	2.73	Domestic
53	Audi 5000	9,690	17 Excellent		3.0	15	2,830	189	37	131	3.20	Foreign
54	Audi Fox	6,295	23 Average		2.5	11	2,070	174	36	97	3.70	Foreign
55	BMW 320i	9,735	25 Good		2.5	12	2,650	177	34	121	3.64	Foreign
56	Datsun 200	6,229	23 Good		1.5	6	2,370	170	35	119	3.89	Foreign
57	Datsun 210	4,589	35 Excellent		2.0	8	2,020	165	32	85	3.70	Foreign
58	Datsun 510	5,079	24 Good		2.5	8	2,280	170	34	119	3.54	Foreign
59	Datsun 810	8,129	21 Good		2.5	8	2,750	184	38	146	3.55	Foreign
60	Fiat Strada	4,296	21 Average		2.5	16	2,130	161	36	105	3.37	Foreign
61	Honda Accord	5,799	25 Excellent		3.0	10	2,240	172	36	107	3.05	Foreign
62	Honda Civic	4,499	28 Good		2.5	5	1,760	149	34	91	3.30	Foreign

Variables

 Filter variables here

<input checked="" type="checkbox"/> Name	Label	Type	Format
<input checked="" type="checkbox"/> price	Price	int	%8.0gc
<input checked="" type="checkbox"/> mpg	Mileage (mpg)	int	%8.0g
<input checked="" type="checkbox"/> rep78	Repair record 1978	int	%9.0g
<input checked="" type="checkbox"/> headroom	Headroom (in.)	float	%6.1f
<input checked="" type="checkbox"/> trunk	Trunk space (cu. ft.)	int	%8.0g
<input checked="" type="checkbox"/> weight	Weight (lbs.)	int	%8.0gc
<input checked="" type="checkbox"/> length	Length (in.)	int	%8.0g
<input checked="" type="checkbox"/> turn	Turn circle (ft.)	int	%8.0g
<input checked="" type="checkbox"/> displacement	Displacement (cu. in.)	int	%8.0g
<input checked="" type="checkbox"/> gear_ratio	Gear ratio	float	%6.2f
<input checked="" type="checkbox"/> foreign	Car origin	byte	%8.0g

Variables			
Filter variables here			
<input checked="" type="checkbox"/>	Name	Label	Type Format
<input checked="" type="checkbox"/>	price	Price	int %8.0gc
<input checked="" type="checkbox"/>	mpg	Mileage (mpg)	int %8.0g
<input checked="" type="checkbox"/>	rep78	Repair record 1978	int %9.0g
<input checked="" type="checkbox"/>	headroom	Headroom (in.)	float %6.1f
<input checked="" type="checkbox"/>	trunk	Trunk space (cu. ft.)	int %8.0g
<input checked="" type="checkbox"/>	weight	Weight (lbs.)	int %8.0gc
<input checked="" type="checkbox"/>	length	Length (in.)	int %8.0g
<input checked="" type="checkbox"/>	turn	Turn circle (ft.)	int %8.0g
<input checked="" type="checkbox"/>	displacement	Displacement (cu. in.)	int %8.0g
<input checked="" type="checkbox"/>	gear_ratio	Gear ratio	float %6.2f
<input checked="" type="checkbox"/>	foreign	Car origin	byte %8.0g

Figure 13: Different types of data.

- Typical storage types are:
  - Byte (integer values between -127 and 100)
  - Int (integer values between -32,767 and 32,740)
  - Long (integer values between -2,147,483,647 and 2,147,483,620)
  - Float (numbers with decimal places with about 8 digits of accuracy)
  - Double (numbers with decimal places with about 16 digits of accuracy)

○ You are now ready to switch to Stata! Let's create the "**stata-class.do**" do-file.

# Data analysis with Stata

## Descriptive statistics

In a report (essay, paper...) it is always useful to start your analysis by providing the reader with some descriptive statistics of the data. For instance, from the data we used so far, we can say that the dataset includes 74 observations of 52 automobiles with US origin and 22 automobiles with foreign origin, and that the sample displays a mean price of about 6,165 dollars, ranging from a minimum of 3,291 to a maximum of 15,906, *etc. etc...* You can include tables (like the ones reproduced below) and a description of them. Be careful: tables should be self-explanatory, so you may want to add some information in the “Table notes”, as in the below example.

Table 1: Descriptive statistics of the main variables.

Variable	Obs.	Mean	Std. Dev.	Min	Max
price	72	5914.208	2562.398	3291	13594
mpg	72	21.403	5.801	12	41
rep	67	3.433	.988	1	5
weight	72	2989.583	765.822	1760	4840

*Table notes:* Price is measured in dollars. Mpg represents the mileage (miles). Rep represents the repair record in 1978, and is coded as 1 “Poor”, 2 “Fair”, 3 “Average”, 4 “Good”, 5 “Excellent”. Weight is measured in Punds (lb).

Sometimes it is useful to display the same statistics of the same variables but for different groups. For instance, in Table 2 we split between domestic cars and foreign cars. We can see that the mean price of the foreign cars is around 6,400 \$. Also, we can see that the mean price of the domestic cars is about 5,700 \$. *Etc. etc...*

Table 2: Descriptive statistics of the main variables, by origin.

Car origin: Domestic					
	N	Mean	SD	Min	Max
price	50	5707.2	2534.666	3291	13594
mpg	50	19.92	4.763	12	34
rep	46	3.043	0.842	1	5
weight	50	3286	689.948	1800	4840
Car origin: Foreign					
price	22	6384.682	2621.915	3748	12990
mpg	22	24.773	6.611	14	41
rep	21	4.286	0.717	3	5
weight	22	2315.909	433.003	1760	3420

*Table notes:* ...



## Graphs

It is also always recommendable to provide visual representations of the data and/or of the information that you deem important for your purposes. Thus, you can include graphs, while describing them in the text.

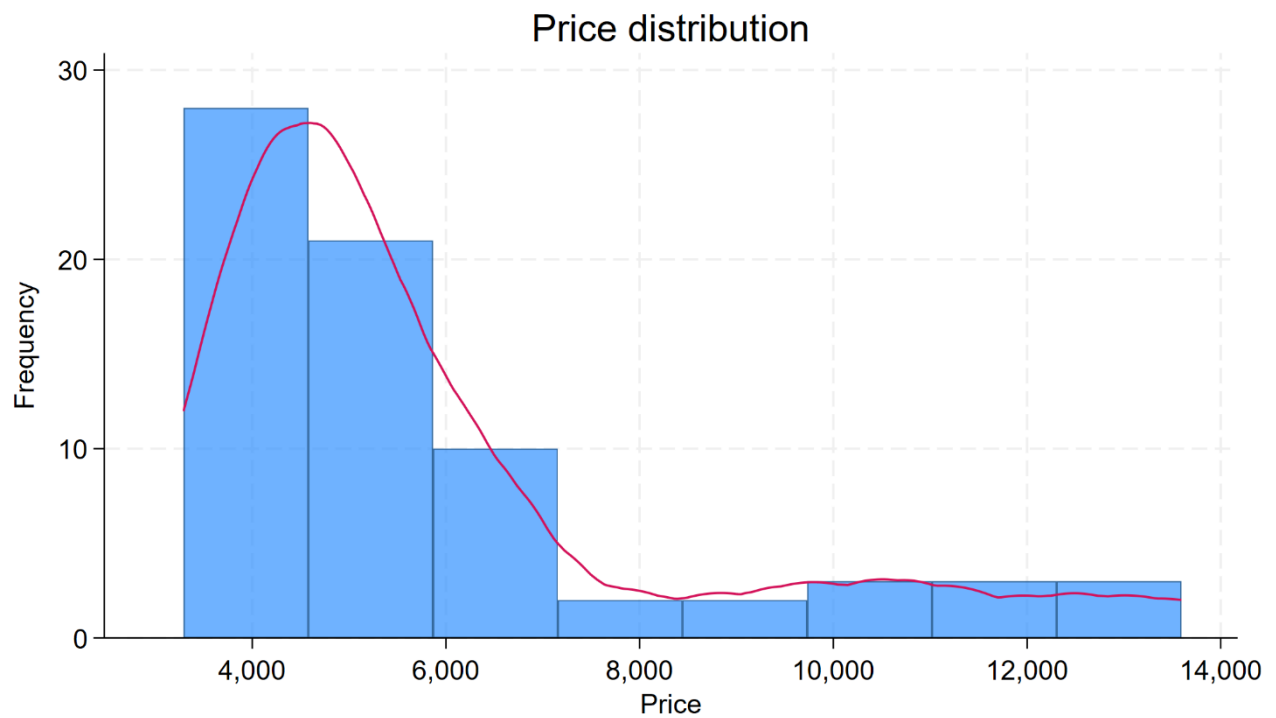


Figure 1: Price distribution.

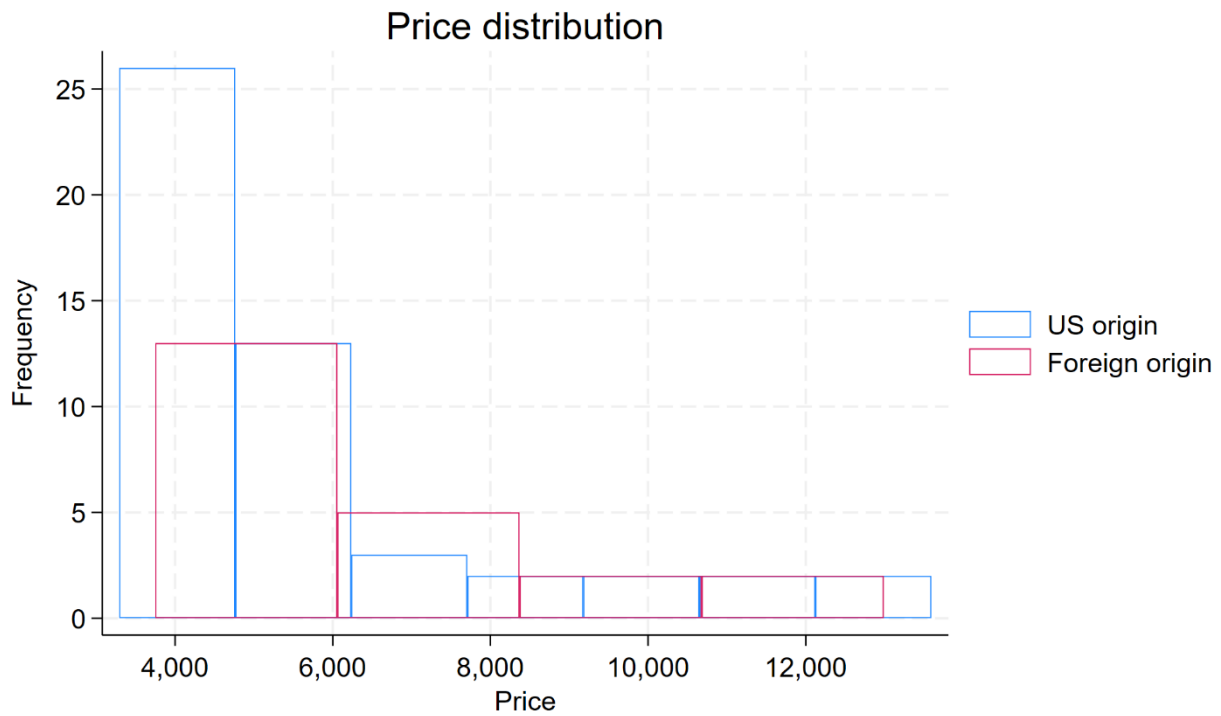


Figure 2: Price distribution by origin.

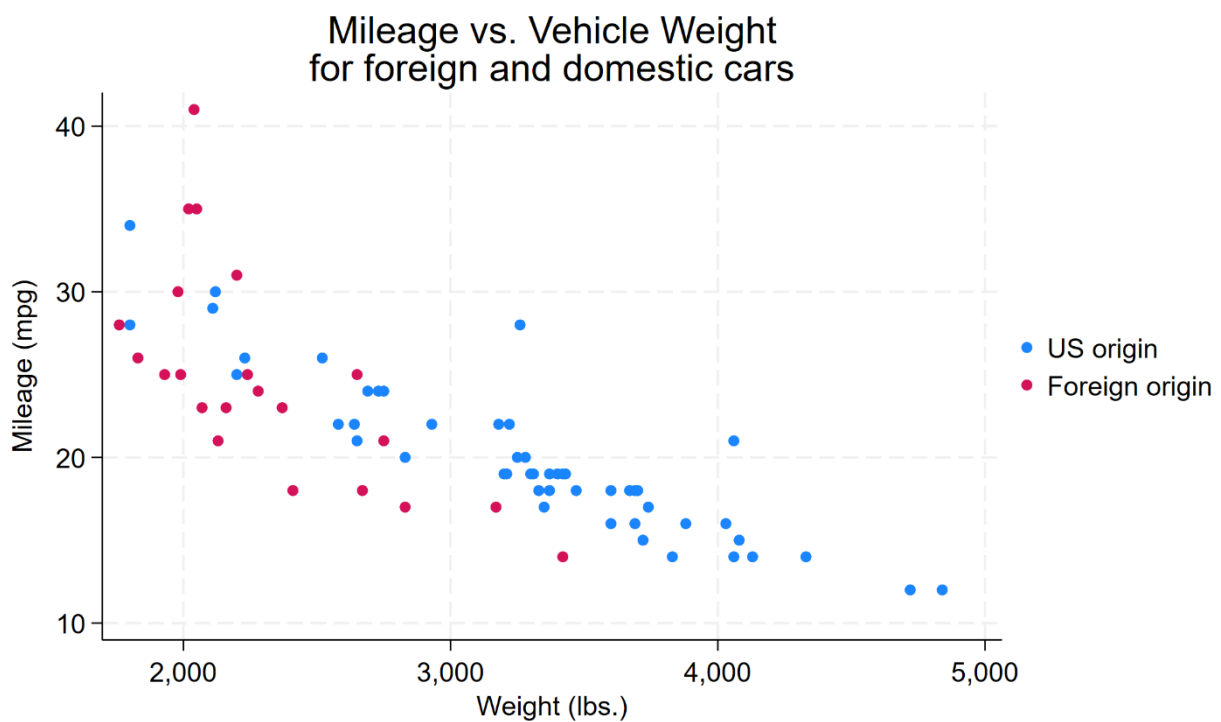


Figure 3: Mileage and Weight relationship.

You can add information on the Pearson Correlation coefficient when commenting a scatter plot. For instance, while describing the trend of Figure 3, we can say that the Pearson Correlation coefficient

between the “mpg” and the “weight” variables is strong and negative (-0.8177) and that, since the relative p-value is less than 0.10, the correlation between these two variables is statistically significant.<sup>1</sup>

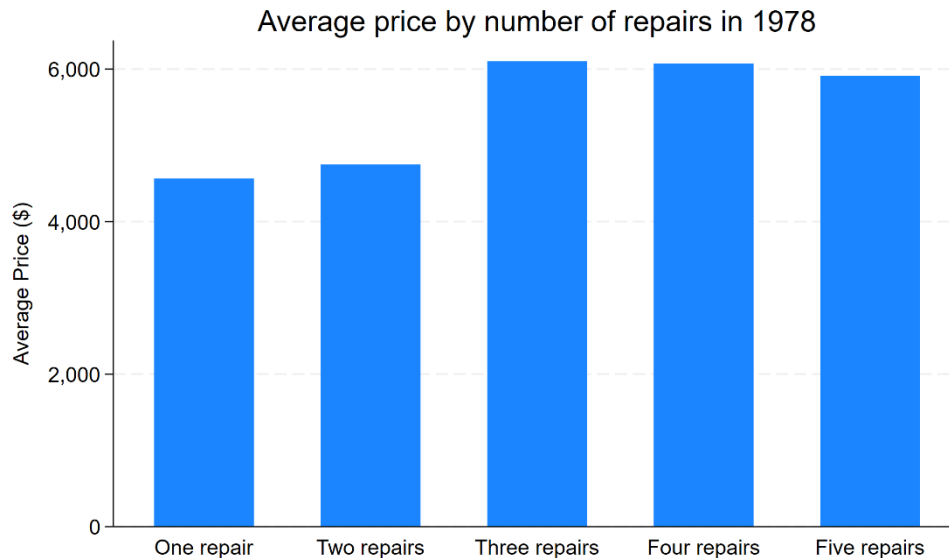


Figure 4: Average price by number of repairs.

This is an example of description for Figure 4:

From the bar graph above, we can see that, on average, prices seem to increase as the status of the car in terms of repair gets better. If the car received only one or two repairs (i.e., it is classified as “Poor” or “Fair”), the mean price is between 4,500 and 5,000 \$; while if the car received at least three repairs (i.e., it is either “Average”, “Good”, or “Excellent”) the mean price is around 6,000 \$ or slightly above.

There is another type of plot that it might be useful (also in your assignments!!!).

It is a time-series line plot. It draws **line plots for time-series data** (for instance, if you have to display the variable over rounds in a given game... ).It can be launched through the command *tsline*. Type “help tsline” in the Command window to know more about it!

---

<sup>1</sup> Note: The significance level used in this document as benchmark is 0.10. However, the standard in the academic community is slightly stricter: usually 0.10 is considered weakly significant; 0.05 is significant; 0.01 is strictly significant.

## Hypothesis testing

When you perform a hypothesis test on a population (in our case, the sample of vintage cars), a **p-value** helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim. This claim is called the null hypothesis ( $H_0$ ). An alternative hypothesis ( $H_a$ ) is the one you would believe correct if the null hypothesis is concluded to be rejected.

The p-value is a probability, i.e., a number between 0 and 1. It can be interpreted in the following way:

- A small p-value (in our case  $\leq 0.10$ ) indicates evidence **against** the null hypothesis. So, we can **reject the null hypothesis**.
- A large p-value ( $> 0.10$ ) indicates that there is **not enough evidence against** the null hypothesis, so we fail to reject the null hypothesis. **We accept the null hypothesis**.

We are going to use a specific type of test called the t-test. You can use it in three ways:

### 1) One sample t-test (a variable equal to a specific value)

```
. ttest mpg == 30
```

One-sample t test

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
mpg	74	21.2973	.6725511	5.785503	19.9569	22.63769

mean = mean(mpg)	t = -12.9398
H0: mean = 30	Degrees of freedom = 73
Ha: mean < 30	Pr(T < t) = 0.0000
Ha: mean != 30	Pr( T  >  t ) = 0.0000
Ha: mean > 30	Pr(T > t) = 1.0000

P-values in Stata are expressed by the “Pr (...)” symbol.

We can accept our null hypothesis (the one written in the command, and that appears labelled as “H0”) if none of the p-values is lower than 0.10. If the “central” p-value is lower than 0.10, we have to reject our null hypothesis.

The “external” p-values give us the direction of the inequality (if “mpg” is different from 30, we want to know whether it is bigger or smaller than 30). We will find our solution, again, with the p-value that is lower than 0.10.

In this case, the mean of “mpg” is lower than 30.

Hence, we can say that: “In this case, we reject the null hypothesis. The mean of mpg is statistically different from 30. More precisely, the mean of mpg is statistically lower than 30.”

## 2) Paired t-test (two variables being equal)

```
. ttest price_1 == price_2
```

Paired t test

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
price_1	74	.4980805	.0322738	.2776299	.4337589	.5624022
price_2	74	.4837835	.0350746	.3017227	.41388	.553687
diff	74	.014297	.0528334	.4544901	-.0909998	.1195938

```
mean(diff) = mean(price_1 - price_2) t = 0.2706
H0: mean(diff) = 0 Degrees of freedom = 73
```

Ha: mean(diff) < 0  
Pr(T < t) = 0.6063

Ha: mean(diff) != 0  
Pr(|T| > |t|) = 0.7875

Ha: mean(diff) > 0  
Pr(T > t) = 0.3937

Our null hypothesis is that the mean of price\_1 is equal to the mean of price\_2. Or, put differently, that the mean of the difference between the two variables is equal to zero. This is why Stata shows us also the hypothesis written in this way: **H0: mean(diff) = 0**.

Let us read the “central” p-value. It is **not** lower than 0.10 (in fact, it is equal to 0.7875). We accept our null hypothesis: the means of the two variables (“price\_1” and “price\_2”) are statistically equal. In this case, hence, we do not need to look at the “external” p-values.

In this case, the means of price\_1 and price\_2 are statistically equal.

Hence, we can say that: **“In this case, we accept the null hypothesis. The mean of price\_1 is statistically equal to the mean of price\_2.”**

## 3) Two sample t-test (a variable being equal across different groups)

```
. ttest mpg, by(foreign)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
Domestic	52	19.82692	.657777	4.743297	18.50638	21.14747
Foreign	22	24.77273	1.40951	6.611187	21.84149	27.70396
Combined	74	21.2973	.6725511	5.785503	19.9569	22.63769
diff		-4.945804	1.362162		-7.661225	-2.230384

```
diff = mean(Domestic) - mean(Foreign) t = -3.6308
H0: diff = 0 Degrees of freedom = 72
```

Ha: diff < 0  
Pr(T < t) = 0.0003

Ha: diff != 0  
Pr(|T| > |t|) = 0.0005

Ha: diff > 0  
Pr(T > t) = 0.9997

Here, we are testing if the mean of a variable is equal between two groups. In this case, we are testing if the mean of "mpg" is equal between foreign cars vs. domestic cars (or, alternatively, if the difference of the two means is equal to zero).

This last test has to be interpreted following the same rules of the previous two tests.

As always, let's first look at the "central" p-value. This is lower than 0.10. We can reject our null hypothesis. This means that the mean of "mpg" is different across domestic and foreign cars.

We look at the "external" p-values and we notice that the difference between the means is lower than 0. If the difference is lower than 0, it means that the first mean is smaller than the second mean. So, the mean of "mpg" for domestic cars is smaller than the mean of "mpg" for foreign cars. Or, in other words, "mpg" is, on average, lower in the domestic cars group. Equivalently, "mpg" is, on average, higher in the foreign cars group.

Hence, we can state that: **"We reject the null hypothesis of "mpg" being on average equal across domestic and foreign cars. The variable "mpg" is statistically different between the two groups. More specifically, the mean of "mpg" is higher among foreign cars than domestic cars".**

## OLS regressions (Advanced -- NOT required in the assignments)

If we want to investigate further the relationship between two variables, say the weight of a car and its miles per gallon, we can use *regressions*. For instance, we can perform a linear regression using weight as an explanatory variable and mpg as an outcome variable, and write the below formula:

$$mpg_i = \alpha + \beta_1 weight_i + \varepsilon_i$$

In Stata, you can write this formula with the command: `regress mpg weight`

<b>. regress mpg weight</b>						
Source	SS	df	MS	Number of obs	=	74
Model	1591.9902	1	1591.9902	F(1, 72)	=	134.62
Residual	851.469256	72	11.8259619	Prob > F	=	0.0000
				R-squared	=	0.6515
				Adj R-squared	=	0.6467
Total	2443.45946	73	33.4720474	Root MSE	=	3.4389
mpg	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
weight	-.0060087	.0005179	-11.60	0.000	-.0070411	-.0049763
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

**How to read this Stata output?** These result show that there is a statistically significant relationship between weight and mpg ( $t = -11.60$ ,  $p < 0.001$ ) and that weight accounts for 65.15% of explained variability in mpg (look at the “R-squared”). On average, each additional pound is associated with a decrease of -0.006 miles per gallon (look at the coefficient attached to the “weight” variable).

Now, we can add some “control variables”. For instance, we add “length”, “trunk”, and “gear\_ratio”. This is a way to reduce residual variance and hence increase precision of the estimate.

$$mpg_i = \alpha + \beta_1 weight_i + \beta_2 length_i + \beta_3 trunk_i + \beta_4 gear\_ratio_i + \varepsilon_i$$

```
. regress mpg weight length trunk gear_ratio
```

Source	SS	df	MS	Number of obs	=	74
Model	1617.27522	4	404.318806	F(4, 69)	=	33.77
Residual	826.184236	69	11.9736846	Prob > F	=	0.0000
				R-squared	=	0.6619
				Adj R-squared	=	0.6423
Total	2443.45946	73	33.4720474	Root MSE	=	3.4603

mpg	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
weight	-.0037126	.0017879	-2.08	0.042	-.0072794	-.0001458
length	-.0756712	.0609814	-1.24	0.219	-.1973257	.0459832
trunk	-.0311581	.1382319	-0.23	0.822	-.3069232	.2446069
gear_ratio	.2914522	1.372298	0.21	0.832	-2.446208	3.029112
_cons	46.27835	8.383549	5.52	0.000	29.55362	63.00307

Our result (the negative and significant coefficient attached to the weight variable) holds also in this second model.

Put together, the two models' results look like the following table. This is the nice output we get by using the command "outreg2" and combining together all the different models.

VARIABLES	(1) Model 1	(2) Model 2
weight	-0.00619*** (0.000521)	-0.00459** (0.00185)
length		-0.0522 (0.0626)
trunk		-0.00454 (0.137)
gear_ratio		0.307 (1.377)
Constant	39.92*** (1.608)	44.05*** (8.616)
Observations	72	72
R-squared	0.669	0.673

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1