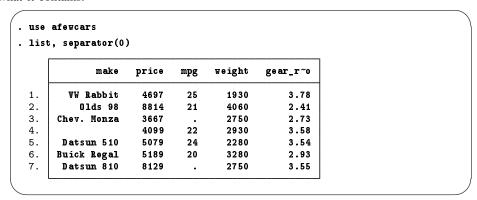
9 Labeling data

Making data readable

This chapter discusses, in brief, labeling of the dataset, variables, and values. Such labeling is critical to careful use of data. Labeling variables with descriptive names clarifies their meanings. Labeling values of numerical categorical variables ensures that the real-world meanings of the encodings are not forgotten. These points are crucial when sharing data with others, including your future self. Labels are also used in the output of most Stata commands, so proper labeling of the dataset will produce much more readable results. We will work through an example of properly labeling a dataset, its variables, and the values of one encoded variable.

The dataset structure: The describe command

At the end of *The import delimited command* in [GSW] **8 Importing data**, we saved a dataset called afewcars.dta. We will put this dataset into a shape that a colleague would understand. Let's see what it contains.



The data allow us to make some guesses at the values in the dataset, but, for example, we do not know the units in which the price or weight is measured, and the term "mpg" could be confusing for people outside the United States. Perhaps we can learn something from the description of the dataset. Stata has the aptly named describe command for this purpose (as we saw in [GSW] 1 Introducing Stata—sample session).

Contains data from afe		cars.dta		
Observations: Variables:		7 5		29 Mar 2023 09:11
Variable	Storage	Display	Value	
name	type	format	label	Variable label
make	str18	%18s		
price	float	%9.0g		
mpg	float	%9.0g		
weight	float	%9.0g		
gear_ratio	float	%9.0g		

Though there is precious little information that could help us as a researcher, we can glean some information here about how Stata thinks of the data from the first three columns of the output.

- 1. The Variable name is the name we use to tell Stata about a variable.
- 2. The *Storage type* (otherwise known as the *data type*) is the way in which Stata stores the data in a variable. There are six different storage types, each having its own memory requirement:
 - a. For integers:

byte for integers between -127 and 100 (using 1 byte of memory per observation) int for integers between -32,767 and 32,740 (using 2 bytes of memory per observation) long for integers between -2,147,483,647 and 2,147,483,620 (using 4 bytes of memory per observation)

b. For real numbers:

float for real numbers with 8.5 digits of precision (using 4 bytes of memory per observation)

double for real numbers with 16.5 digits of precision (using 8 bytes of memory per observation)

c. For strings (text) between 1 and 2,045 bytes (using 1 byte of memory per observation per character for ASCII and up to 4 bytes of memory per Unicode character):

```
str1 for 1-byte-long strings
str2 for 2-byte-long strings
str3 for 3-byte-long strings
...
str2045 for 2,045-byte-long strings
```

d. Stata also has a strL storage type for strings of arbitrary length up to 2,000,000,000 bytes. strLs can also hold binary data, often referred to as BLOBs, or binary large objects, in databases. We will not illustrate these here.

Storage types affect both the precision of computations and the size of datasets. A quick guide to storage types is available at help data types or in [D] **Data types**.

3. The *Display format* controls how the variable is displayed; see [U] 12.5 Formats: Controlling how data are displayed. By default, Stata sets it to something reasonable given the storage type.

We would like to make this dataset into something containing all the information we need.

To see what a well-labeled dataset looks like, we can look at a dataset stored at the Stata Press repository. We need not load the data (and disturb what we are doing); we do not even need a copy of the dataset on our machine. (You will learn more about Stata's Internet capabilities in [GSW] 19 Updating and extending Stata—Internet functionality.) All we need to do is direct describe to look at the proper file by using the command describe using filename.

Contains data Observations Variables	:	74 12		1978 automobile data 13 Apr 2022 17:45
Variable name	Storage type	Display format	Value 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
he adro om	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 origin

This output is much more informative. There are three locations where labels are attached that help explain what the dataset contains:

- 1. In the first line, 1978 automobile data is the data label. It gives information about the contents of the dataset. Data can be labeled by selecting Data > Data utilities > Label utilities > Label dataset, by using the label data command, or by editing the Label field in the Data portion of the Properties window. When doing this in the main window, be sure that the Properties window is unlocked.
- 2. There is a variable label attached to each variable. Variable labels are how we would refer to the variable in normal, everyday conversation. Here they also contain information about the units of the variables. Variables can be labeled by selecting the variable in the Variables window and editing the Label field in the Properties window. You can also change a variable label by using the Variables Manager or by using the label variable command.
- 3. The foreign variable has an attached value label. Value labels allow numeric variables such as foreign to have words associated with numeric codes. The describe output tells you that the numeric variable foreign has value label origin associated with it. Although not revealed by describe, the variable foreign takes on the values 0 and 1, and the value label origin associates 0 with Domestic and 1 with Foreign. If you browse the data (see [GSW] 6 Using the Data Editor), foreign appears to contain the values "Domestic" and "Foreign". The values in a variable are labeled in two stages. The value label must first be defined. This can be done in the Data Editor, or in the Variables Manager, or by selecting Data > Data utilities > Label utilities > Manage value labels or by typing the label define command. After the labels have been defined, they must be attached to the proper variables, either by selecting **Data > Data utilities** > Label utilities > Assign value label to variables or by using the label values command.

Note: It is not necessary for the value label to have a name different from that of the variable. You could just as easily have used a value label named foreign.

Labeling datasets and variables

We will now load the afewcars.dta dataset and give it proper labels. We will do this with the Command window to illustrate that it is simple to do in this fashion. Earlier in *Renaming and formatting variables* in [GSW] 6 Using the Data Editor, we used the Data Editor to achieve a similar purpose. If you use the Data Editor for the material here, you will end up with the same commands in your log; we would like to illustrate a way to work directly with commands.

describe				
ontains dat				
Observation		7		
Variable	es:	5		29 Mar 2023 09:11
/ariable	Storage	Display	Value	
name	type	format	label	Variable label
nake	str18	%18s		
rice	float	%9.0g		
npg	float	%9.0g		
reight	float	%9.0g		
gear_ratio	float	%9.0g		
Sorted by:	ı "A few 19	78 cars"		
•	able make able price able mpg "! able weigh able gear_: a from afer	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear	D)" les per ga weight (lb	
label data label vari label vari label vari label vari describe Contains dat Observation	able make able price able mpg "! able weigh able gear_: a from afer	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear wcars.dta 7	D)" les per ga weight (lb	s.)" A few 1978 cars
label data label vari label vari label vari label vari label vari describe Contains dat Observation Variable	able make able price able mpg "! able weigh able gear_: a from afer s:	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear wcars.dta 7 5	D)" les per ga weight (lb ratio"	s.)" A few 1978 cars
label data label vari label vari label vari label vari label vari describe Contains dat Observation Variable	able make able price able mpg "! able weigh able gear_: a from afer s: Storage	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear wcars.dta 7 5 Display	D)" les per ga weight (lb ratio"	S.)" A few 1978 cars 29 Mar 2023 09:11
label data label vari label vari label vari label vari describe contains dat Observation Variable fariable mame	able make able price able mpg "! able weigh able gear_; a from afer as: Storage type	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear wcars.dta 7 5 Display format	D)" les per ga weight (lb ratio"	A few 1978 cars 29 Mar 2023 09:11 Variable label
label data label vari label vari label vari label vari label vari describe Contains dat Observation Variable Variable	able make able price able mpg "! able weigh able gear_; a from afer as: ss: Storage type str18	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear Wcars.dta 7 5 Display format %18s	D)" les per ga weight (lb ratio"	A few 1978 cars 29 Mar 2023 09:11 Variable label Nake and model
label data label vari label vari label vari label vari describe contains dat Observation Variable fariable mame	able make able price able mpg "! able weigh able gear; a from afer as: ss: Storage type str18 float	"Make and m "Price (US Mileage (mi t "Vehicle ratio "Gear Gear Gear T Display format %18s %9.0g	D)" les per ga weight (lb ratio"	A few 1978 cars 29 Mar 2023 09:11 Variable label Make and model Price (USD)

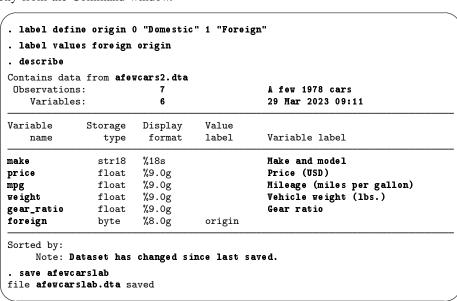
Labeling values of variables

We will now add a new indicator variable to the dataset that is 0 if the car was made in the United States and 1 if it was made in another country. Open the Data Editor and use your previously gained knowledge to add a foreign variable whose values match what is shown in this listing:

	-					
	make	price	mpg	weight	gear_r~o	foreign
l	VW Rabbit	4697	25	1930	3.78	1
İ	01ds 98	8814	21	4060	2.41	0
ı	Chev. Monza	3667		2750	2.73	0
İ		4099	22	2930	3.58	0
İ	Datsun 510	5079	24	2280	3.54	1
	Buick Regal	5189	20	3280	2.93	0
	Datsun 810	8129		2750	3.55	1

You can create this new variable in the Data Editor if you would like to work along. (See [GSW] 6 Using the Data Editor for help with the Data Editor.) Though the definitions of the categories "0" and "1" are clear in this context, it still would be worthwhile to give the values explicit labels because it will make output clear to people who are not so familiar with antique automobiles. This is done with a value label.

We saw an example of creating and attaching a value label by using the point-and-click interface available in the Data Editor in Changing data in [GSW] 6 Using the Data Editor. Here we will do it directly from the Command window.



From this example, we can see that a value label is defined via

label define labelname # "contents" # "contents" ...

It can then be attached to a variable via

label values variablename labelname

Once again, we need to save the dataset to be sure that we do not mistakenly lose the labels later. We saved this under a new filename because we have cleaned it up, and we would like to use it in the next chapter.

If you had wanted to define the value labels by using a point-and-click interface, you could do this with the Properties window in either the Main window or the Data Editor or by using the Variables Manager. See [GSW] 7 Using the Variables Manager for more information.

There is more to value labels than what was covered here; see [U] 12.6.3 Value labels for a complete treatment.

You may also add notes to your data and your variables. This feature was previously discussed in Renaming and formatting variables in [GSW] 6 Using the Data Editor and Managing notes in [GSW] 7 Using the Variables Manager. You can learn more about notes by typing help notes, or you can get the full story in [D] notes.

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow and NetCourseNow are trademarks of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985-2023 StataCorp LLC, College Station, TX, USA. All rights reserved.



For suggested citations, see the FAQ on citing Stata documentation.