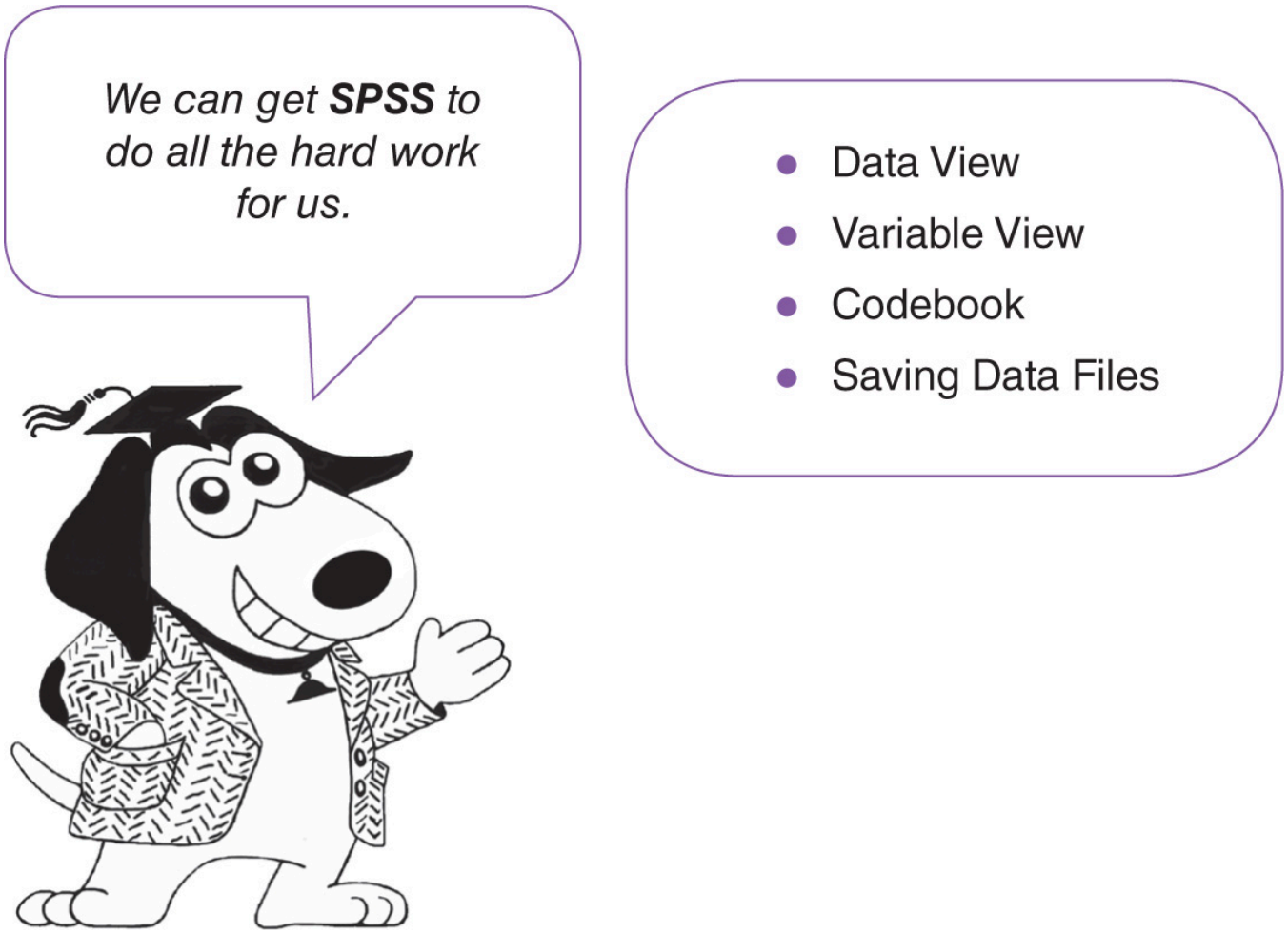
Sage Research Methods

Introductory Statistics Using SPSS

For the most optimal reading experience we recommend using our website. A free-to-view version of this content is available by clicking on this link, which includes an easy-to-navigate-and-search-entry, and may also include videos, embedded datasets, downloadable datasets, interactive questions, audio content, and downloadable tables and resources.

Author: Herschel Knapp
Pub. Date: 2022
Product: Sage Research Methods
DOI: https://doi.org/10.4135/9781071878910
Methods: T-test, Descriptive statistics, Normal distribution
Keywords: screens, property, screening, labeling
Disciplines: Psychology, Mathematics
Access Date: July 19, 2024
Publisher: SAGE Publications, Inc
City: Thousand Oaks
Online ISBN: 9781071878910

Working in SPSS



Computers are useless. They can only give you answers.

—Pablo Picasso

Learning Objectives

Upon completing this chapter, you will be able to:

- Operate in the two primary views in SPSS: Variable View and Data View
- Establish or modify variable definitions on the Variable View screen: name, type, width, decimals, la-

bel, values, missing, columns, align, measure, and role
- Use the value label icon to alternate between numeric and label (text) displays
- Interpret and use a codebook to configure variables in SPSS
- Enter data into SPSS
- Save and identify SPSS data files

Video

The video for this chapter is **Ch 03 - Working in SPSS.mp4**. This video provides guidance on setting up an SPSS database, entering data, and saving SPSS files.

Overview—SPSS

Based on what you have read thus far, you have probably figured out that when it comes to statistics, larger sample sizes facilitate more robust statistical findings. Appropriately large sample sizes are also important when it comes to gathering a *representative sample*, which helps when it comes to generalizing the findings from your sample to the overall population from which it was drawn (*external validity*). Processing large samples can involve hundreds or even thousands of calculations. For most statistical formulas, the mathematical complexity does not go beyond simple algebra, but such formulas typically involve multiple mathematical op-

erations on each record. Attempting to process such data by hand would be inefficient in two ways: (a) It would be very time-consuming, and (b) accuracy would be compromised. Performing multiple calculations on a large data set is bound to produce some errors along the way. Even if each mathematical operation were correct, the data would be vulnerable to cumulative rounding error. With the advent of affordable, powerful computers and menu-driven statistical programs, it is now possible to accurately perform a variety of statistical analyses in a matter of seconds with relative ease. This chapter will provide you with what you need to know to get started using SPSS. SPSS, which originally stood for *Statistical Program for the Social Sciences*, has gone through some substantial evolution over time; some versions are referred to as PASW, for *Predictive Analytics Software.* For the remainder of the text, the term *SPSS* will be used. Regardless of the name, the SPSS functionality of the statistics covered in this text has remained relatively stable across the evolution of the software.

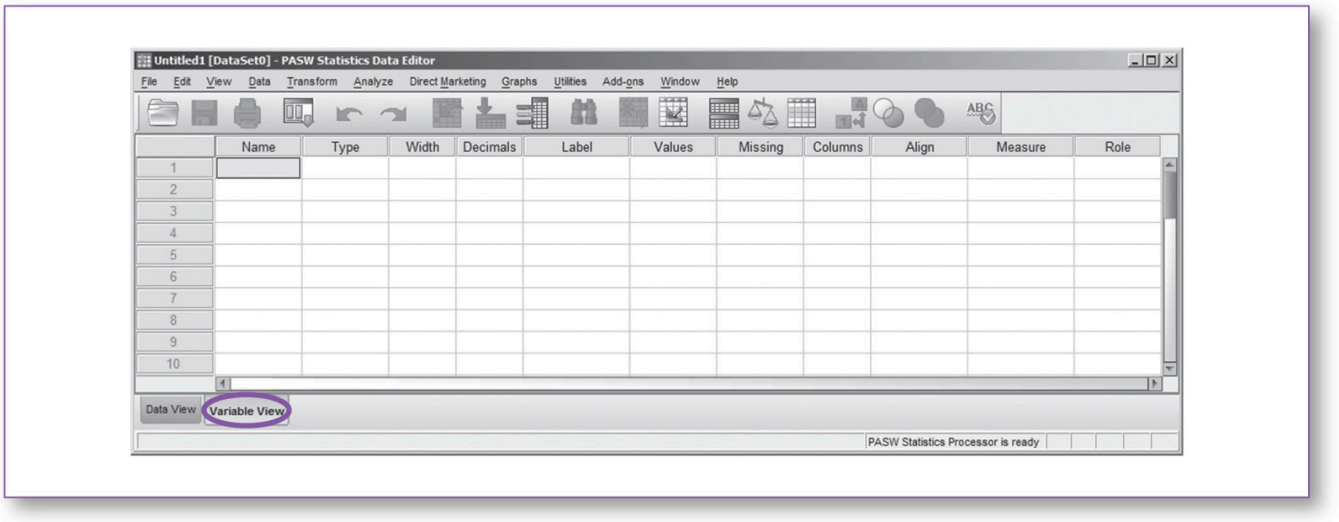
Two Views: Variable View and Data View

SPSS is laid out as two main screens: the *Variable View*, which is used for establishing or modifying the characteristics of each variable, and the *Data View*, which contains the gathered data. We will begin with the *Variable View*.

Variable View

The *Variable View* provides a screen for you to systematically set up the variables that will contain your data. This is where you will assign the name and characteristics of each variable you will be including in the data set. To access the *Variable View* screen, click on the tab at the bottom of the screen that says *Variable View*, as shown in Figure 3.1.

Figure 3.1 SPSS Variable View screen.



Basically, for each variable, you are telling SPSS the name of the variable and the kind of data it will contain (e.g., regular numbers, dates, text), along with some other properties (parameters). Once you have established each variable on the *Variable View* screen, you can proceed to enter the data you have gathered on the *Data View* screen, which resembles a traditional spreadsheet. The *Variable View* screen has 11 properties you can set for each variable. Some versions of SPSS may have a different number of properties; you should be able to proceed nonetheless. Naturally, you will use care when establishing variables on the *Variable View* screen, but there is no need to be nervous; even after you have entered data on the *Data View* screen, you can always return to the *Variable View* screen and make changes (e.g., include more variables, delete variables, rename variables, and modify the properties of existing variables).

The cursor is initially positioned in the *Name* column for the first variable; this is where the data definition process begins.

Name

Each variable needs a unique name. Variable names can contain uppercase and lowercase letters, but variable names are not case sensitive, meaning that you cannot have a variable named *AGE* and another variable (in the same database) named *age* or *Age*. The name can contain up to 64 letters and numbers; the first character must be a letter. Some older versions of SPSS allow only eight characters for variable names, so

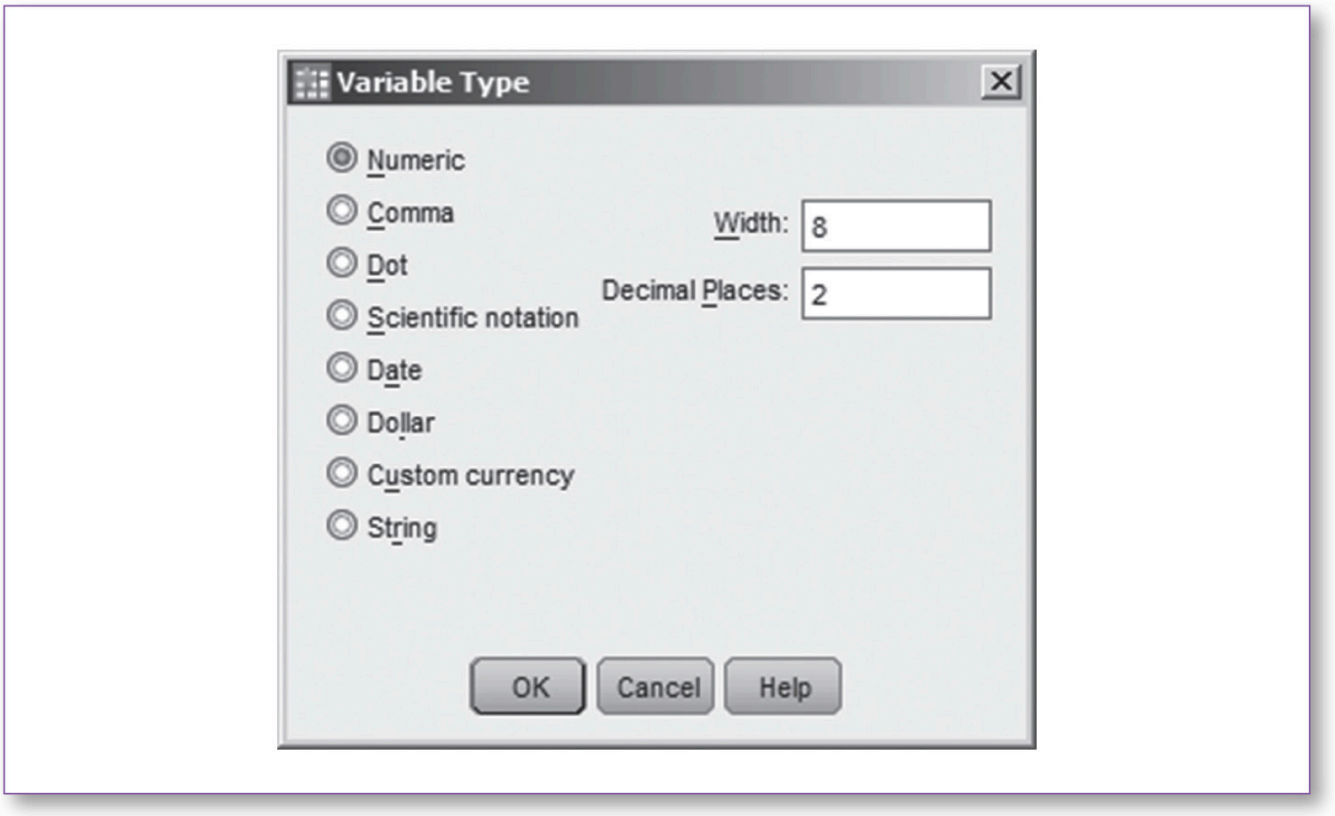
you may need to be imaginative when it comes to assigning briefer variable names. Spaces are not allowed in the variable name, but you can use the underscore (_) character instead. For your own convenience, try to assign meaningful names (e.g., *age*, *date_of_birth*, *first_name*, *last_name*, *gender*, *test01*, *question01*, *question02*, *question03*). It is okay if you are unable to assign a perfect variable name; this will be discussed in more detail when we look at the *Label* property. The cursor is positioned in the first cell for the first variable. We will build a database containing two variables: *gender* (a categorical variable) and *age* (a continuous variable). Begin by entering *gender* (in Row 1, under *Name*). When you press Enter, notice that SPSS automatically enters default values for all of the remaining properties, except for *Label*. Each of these properties can be changed, but we will accept the automatic defaults for some.

Type

The system needs to know what *type* of data the variable will contain. The system assigns the default type as a *numeric* variable with a width of eight integer digits and two decimal digits (which you can change), meaning that this variable will accommodate a number such as 12345678.12 (Figure 3.2).

To access the window shown in Figure 3.2, click on the *Type* cell for that variable. The options for variable type are fairly self-explanatory except for *String*; a *string* variable contains alphanumeric (letters and numbers) data (e.g., name, note, comment, memo). A string variable is useful for data that contain information that will not be processed mathematically, consisting of letters, numbers, punctuation, or a mixture of letters and numbers, such as an ID code, an address, or a name (e.g., APB-373, 852 S. Bedford Street, Dusty Jones); if your variable is not a date or a numeric value, then consider it alphanumeric and select the *String* type. While string variables may contain valuable information, it is not possible to perform statistical operations on such variables.

Figure 3.2 SPSS Variable Type window.



Width

The *Width* refers to the number of characters SPSS will allow you to enter for the variable. If it is a numeric value with decimals, the total *Width* has to include the decimal point and each digit (Figure 3.2). For example, if you were entering data for income (including dollars and cents) and the largest value expected was 123456.78, you would want to be sure to set the *Width* to at least 9 (8 digits for the numbers plus 1 for the decimal point).

Decimals

The *Decimals* property refers to the number of digits that will appear to the right of the decimal point (Figure

3.2). The default is two decimal places. For example, a cash amount should have two decimal places (for the cents), whereas for categorical variables (e.g., gender: Female = 1, Male = 2), a decimal is not needed, so the value should be set to zero.

Label

If the *Label* property is left blank, SPSS will use the variable name in all output reports; otherwise, it will use whatever you specify as the label. For example, suppose the name of the variable is *dob*, but in your reports, you want it to display as “Date of Birth”; in that case, simply enter “Date of Birth” in the *Label* property. Notice that the *Label* can contain spaces, but the *Name* cannot.

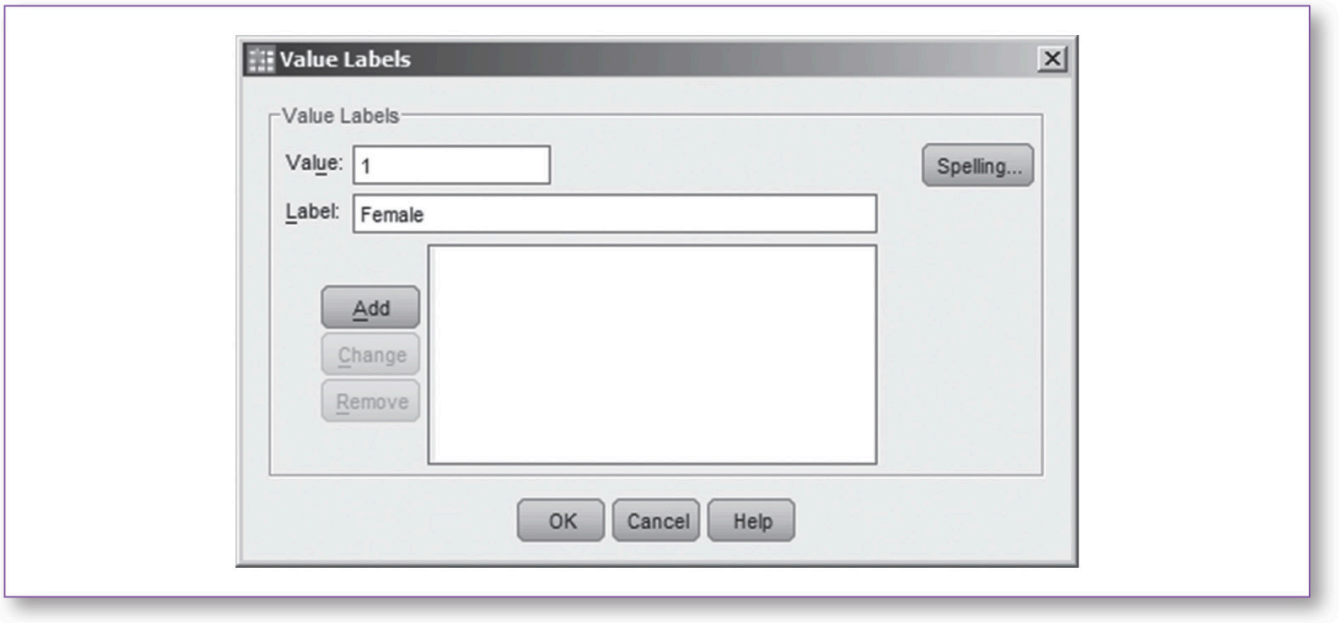
Values

The *Values* property provides a powerful instrument for assigning meaningful names to the values (numbers) contained in categorical variables. For example, *gender* is a nominal variable containing two categories (1 = Female, 2 = Male). When it comes to nominal variables, SPSS handles categories as numbers (1, 2) as opposed to the textual names (Female, Male). The *Values* property allows you to assign the textual name to each category number, so even though you will code *gender* using 1’s and 2’s, the output reports will exhibit these numerals as “Female” and “Male.”

Here is how it works:

1. In the Name column, create a variable called **gender**; accept all the default values, except change the Decimals property to 0.
2. Click on the Values cell for **gender**; this will bring up the Value Labels window (Figure 3.3).
3. Assign the values one at a time; begin by entering 1 in Value and “Female” in Label, then click Add.
4. Do the same for the second category: Enter 2 in Value and “Male” in Label, then click Add.
5. To finalize these designations, click OK.

Figure 3.3 SPSS Value Labels window.



You will see the utility of this labeling system when you enter data on the *Data View* screen and when you run your first report.

Missing

Sometimes, when the source data are either erroneous or missing, the cell is simply left blank, in which case the *Missing* property can remain blank as well. Other times, the erroneous or missing data are represented by special numeric codes; a common convention is to code erroneous data as 888, and missing data are represented as 999—this conveys that a blank cell is not an oversight. Consider the variable *age*; if the data contained an erroneous entry (e.g., “I’m a kid”), or if the entry were left blank, the corresponding 888 or 999 code would radically throw off the statistical (age) calculations. The *Missing* property enables us to specify such codes (888 and 999) that we want SPSS to ignore so that they will not be processed in the statistical calculations.

Here is how it works:

1. Create a variable with the name **age**; accept all the default values, except change the Decimals prop-

erty to 0.

2. Click on Discrete missing values and enter 888 and 999 (Figure 3.4).
3. If you need to indicate more than three such values, you may opt for the *Range plus one optional discrete missing value* function, which would enable you to specify a range of values (e.g., Low: 888, High: 999, meaning that all values from 888 through 999 inclusive will be omitted from all statistical analysis for that variable). In addition, you can specify one additional value (e.g., Discrete value: −1).
4. To finalize these designations, click OK.

The numbers 888 and 999 have been generally adopted as special values since they are visually easy to recognize in a data set. Also, if these special values are not properly designated as erroneous or missing values, statistical clues will begin to emerge, such as a report indicating an average age of 347 or a maximum height of 999 inches or centimeters. Such extreme results alert you to check that the missing and erroneous designations have been properly specified for a variable.

Figure 3.4 SPSS Missing Values window.

Columns

The *Columns* property allows you to change the column width on the *Data View* screen and in the reports; you can specify how many characters wide you want a column to be.

Align

The *Align* property lets you specify how you want the variable to be presented on the *Data View* screen and in the output reports. Typically, *Right* alignment (justification) is used for numeric data and *Left* alignment is used for text, such as string data or categorical variables with data labels assigned to them. *Center* is also an option.

Measure

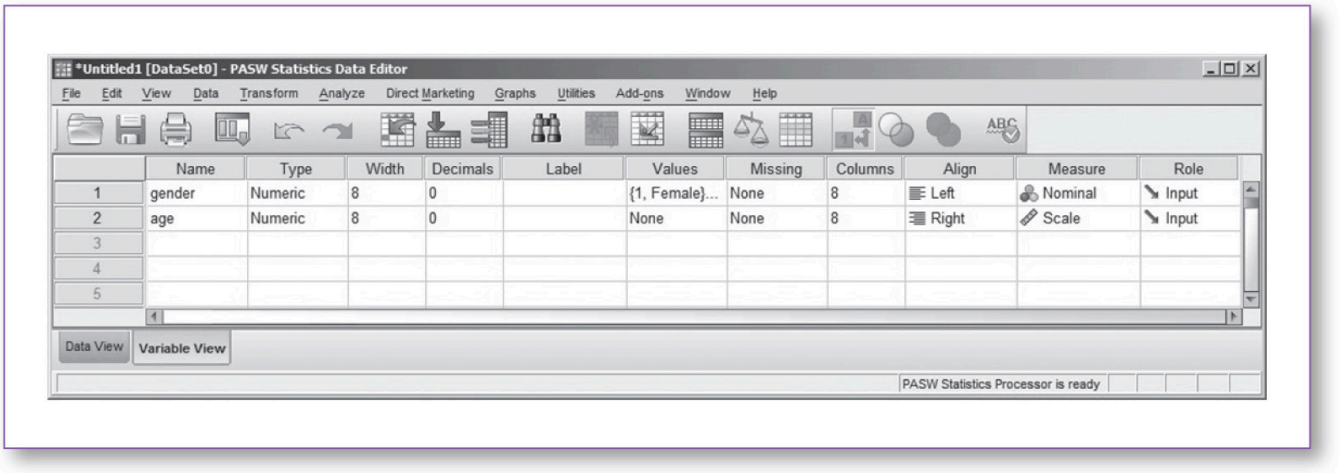
The *Measure* property pertains to the four levels of measures (*nominal*, *ordinal*, *interval*, and *ratio*) covered in the "Variable Types and Levels of Measure" section in Chapter 1. For variables that contain *continuous* (interval or ratio) variables, select *Scale*. For *categorical* (*nominal* or *ordinal*) variables, which may contain value labels, select either *Nominal* or *Ordinal*, depending on the variable type.

Role

Some versions of SPSS have the *Role* property; do not panic if the version you are using does not include this property, as it will not be used in this text. Role enables you to define how the variable will be used in the statistical processes. If your version of the software includes the *Role* property, just use the default setting: *Input*.

Use SPSS to set up the *Variable View* screen to establish the *gender* and *age* variables as shown in Figure 3.5.

Figure 3.5 SPSS Variable View screen.



Data View

Now that the properties for each variable have been established on the *Variable View* screen, the next step is to enter the actual data. To switch to the data entry mode, click on the *Data View* tab at the bottom of the screen. As you enter the data in Table 3.1 into the *Data View* screen, notice that for the *gender* variable, you can access the pull-down menu in each cell to select “Female” or “Male.” Alternatively, you can enter the corresponding numbers that you defined: 1 for female and 2 for male. Notice that SPSS will not allow you to type the words *Female* and *male* directly into the gender field; you will need to enter a number (in this case, 1 or 2) or use the pull-down menu feature to select “Female” or “Male” for this variable. The *Data View* screen should resemble Figure 3.6.

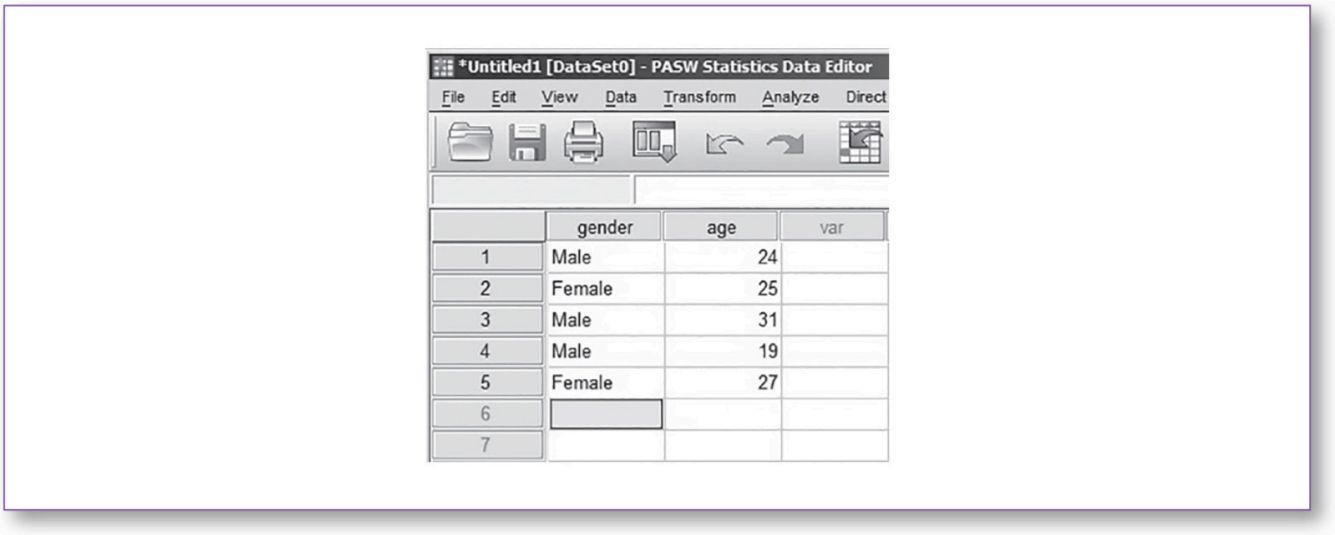
Table 3.1 Source Data for gender and age.

	gender	age
1	Male	24

2	Female	25
3	Male	31
4	Male	19
5	Female	27

NOTE: You do not need to enter the numbers in the shaded leftmost column (1, 2, 3, 4, and 5); this column pertains to the row (record) numbers that SPSS provides automatically.

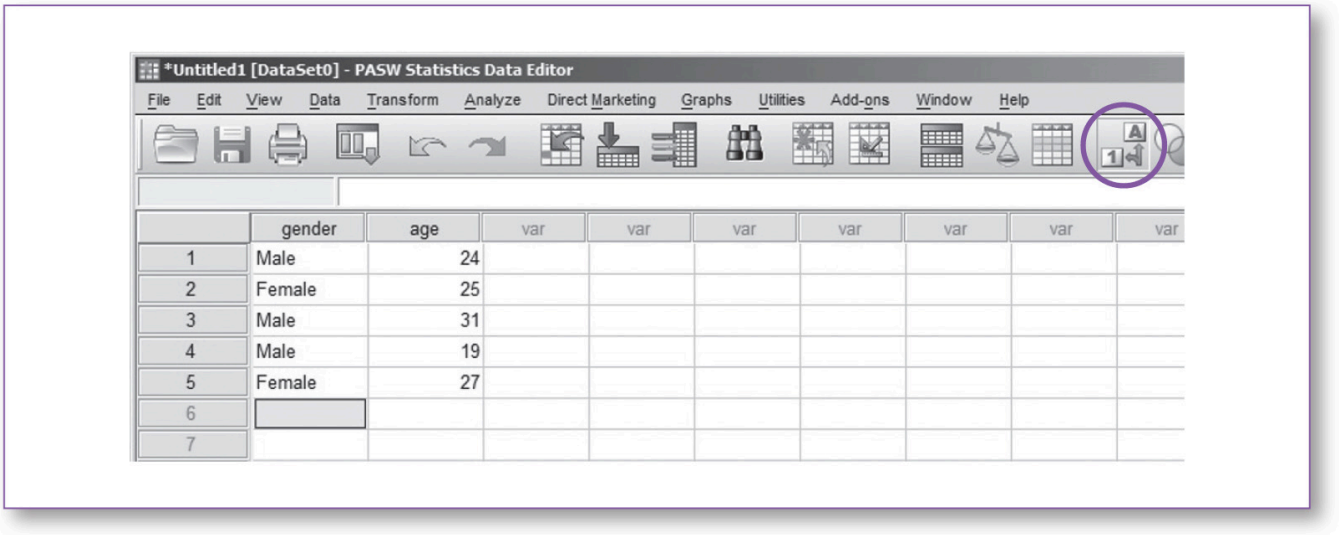
Figure 3.6 Data View screen with data entered.



Value Labels Icon

When it comes to viewing your data, there will be times when you will want to see the value labels (e.g., “Female,” “Male”) and other times when you will want to see the source numbers (e.g., 1, 2). To toggle this display back and forth, from numbers to text (and back), click on the *Value Labels* icon (with the *1* and *A* on it), as shown in Figure 3.7.

Figure 3.7 The Value Labels icon alternates the display of categorical variables from text to numeric display (and back).



Codebook

For learning purposes, the two-variable data set used in this chapter is admittedly simple. Even so, without being told that for the variable *gender*, 1 stands for female and 2 stands for male, this coding scheme would lead to confusion. Designations such as “1 = Female” and “2 = Male” and other characteristics of each variable in a data set are traditionally contained in the *codebook*, which is the companion to the data set. The codebook is written by the person who develops the experiment or survey; it provides a descriptive list of each variable contained in a data set. This is particularly valuable in data sets that contain numerous variables with arcane names. For example, suppose we came across a variable named *Q105* (*Question 105*), and it appeared to contain dates. Without the codebook, we would have no idea what any of this means; we would not know how this variable was gathered, nor would we be able to assign any meaning to these dates (e.g., birth date, death date, graduation date, anniversary, date of arrest, date admitted to a hospital). If you do not know the story of a variable, the data are virtually useless; hence, the codebook is as valuable as the actual data set. Although there is no standard form for codebooks, a quality codebook should indicate the information essential to understanding each variable in the data set. Continuing with the *Q105* example, a reasonable codebook entry for this variable might look like this:

Codebook

Variable: Q105
Definition: High school graduation date (Question #105)
Type: Date (MM/DD/YYYY)

The codebook for our simple two-variable database detailed in Table 3.1 looks like this (this concise codebook format will be used throughout this book):

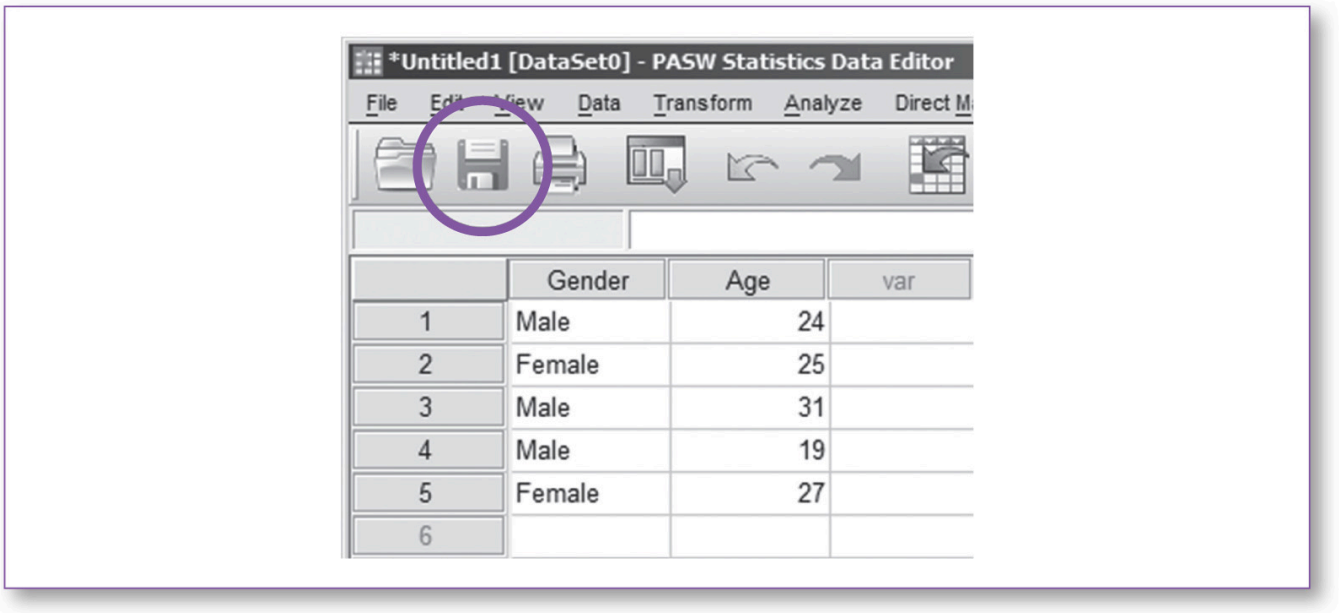
Codebook

Variable: gender
Definition: Gender of respondent
Type: Categorical (1 = Female, 2 = Male)
Variable: age
Definition: Age of respondent
Type: Continuous

Saving Data Files

To save the file, click on the *Save this document* icon, as shown in Figure 3.8. Use the file name "First Data Set"; SPSS automatically appends the *.sav* extension (suffix) to the file name. The file on your system will be listed as *First Data Set.sav*.

Figure 3.8 The Save this document icon.



Good Common Sense

The acronym GIGO (pronounced *gig-oh*) comes from the early days of computing; it stands for "garbage in, garbage out," and it is just as valid today. Basically, it means that if you input inaccurate data into a program, the output will be inaccurate too. Inaccurate data can consist of missing or erroneous responses, responses to misconstrued questions, data entry errors (typos), omitted data, double or multiple entries of data, and other anomalies that may have allowed imprecise data to be entered into the database. Skillful statisticians will inspect and assess the data for such errors prior to embarking on analyses; this process is referred to as *cleaning the data*.

For example, a survey question may ask for the participant's age; however, the respondent enters a date (e.g., "9/18/1980"). Clearly, data in date format will not fit into a variable that is configured to accept a three-digit numeric (age) value. The statistician would then need to make a judgment call: The data could be omitted or coded as 888 (error), or one might presume that the date provided is the participant's birthdate, in which case the age could be calculated and entered into the specified field.

The accuracy of the statistical tests you run will depend on the accuracy of the data definitions (on the *Variable*

View screen) and the data entered (on the *Data View* screen). Considering that this book focuses on learning specific statistics, as opposed to coping with erroneous or missing data, the data sets that are provided contain *clean* data with no missing values, which are ready for processing.

Key Concepts

- Variable View
    - Name
    - Type
    - Width
    - Decimals
    - Label
    - Values
    - Missing
    - Columns
    - Align
    - Measure
    - Role
- Data View
    - Value Labels icon
    - Codebook
    - Saving data files
        - Save this document icon
        - .sav files

Practice Exercises

Use the provided codebook in each exercise to establish the variables on the *Variable View* screen, and then enter the data on the *Data View* screen.

To check your work, produce a variable list; click on *Analyze*, *Reports*, *Codebook*, as shown in Figure 3.9.

Next, select all the variables that you want to include in the codebook report; move the variables from the left *Variables* panel to the right *Codebook Variables* panel (using double-click, drag and drop, or the arrow

button), then click *OK*, as shown in Figure 3.10. This will generate a *Variable Information* report showing the properties of all variables, as shown in Table 3.2.

Figure 3.9 Ordering a list of all variables; click on Analyze, Reports, Codebook.

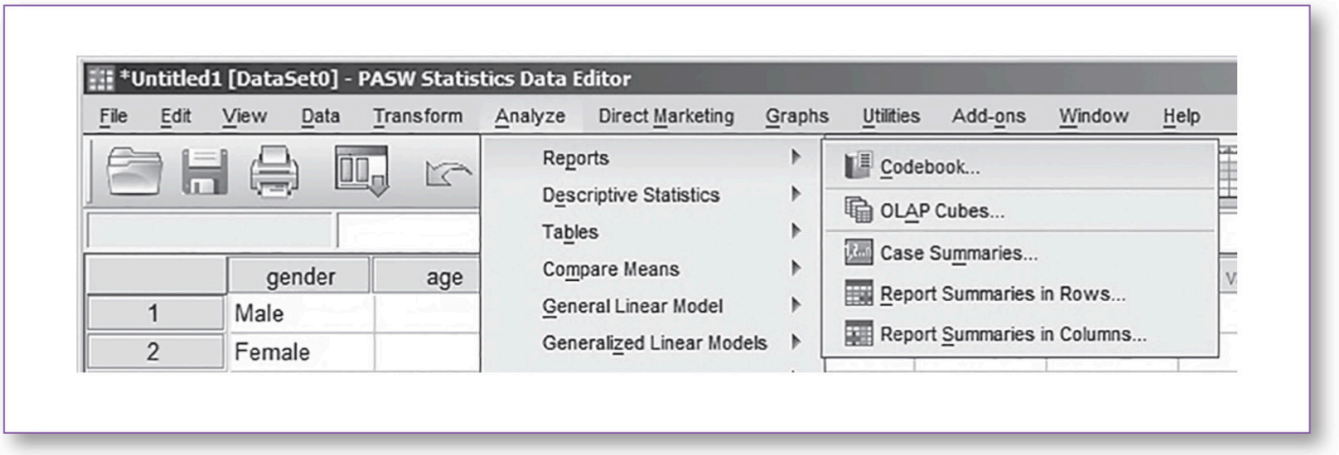


Figure 3.10 Codebook report order screen; move variables of interest to the right (Codebook Variables) panel.

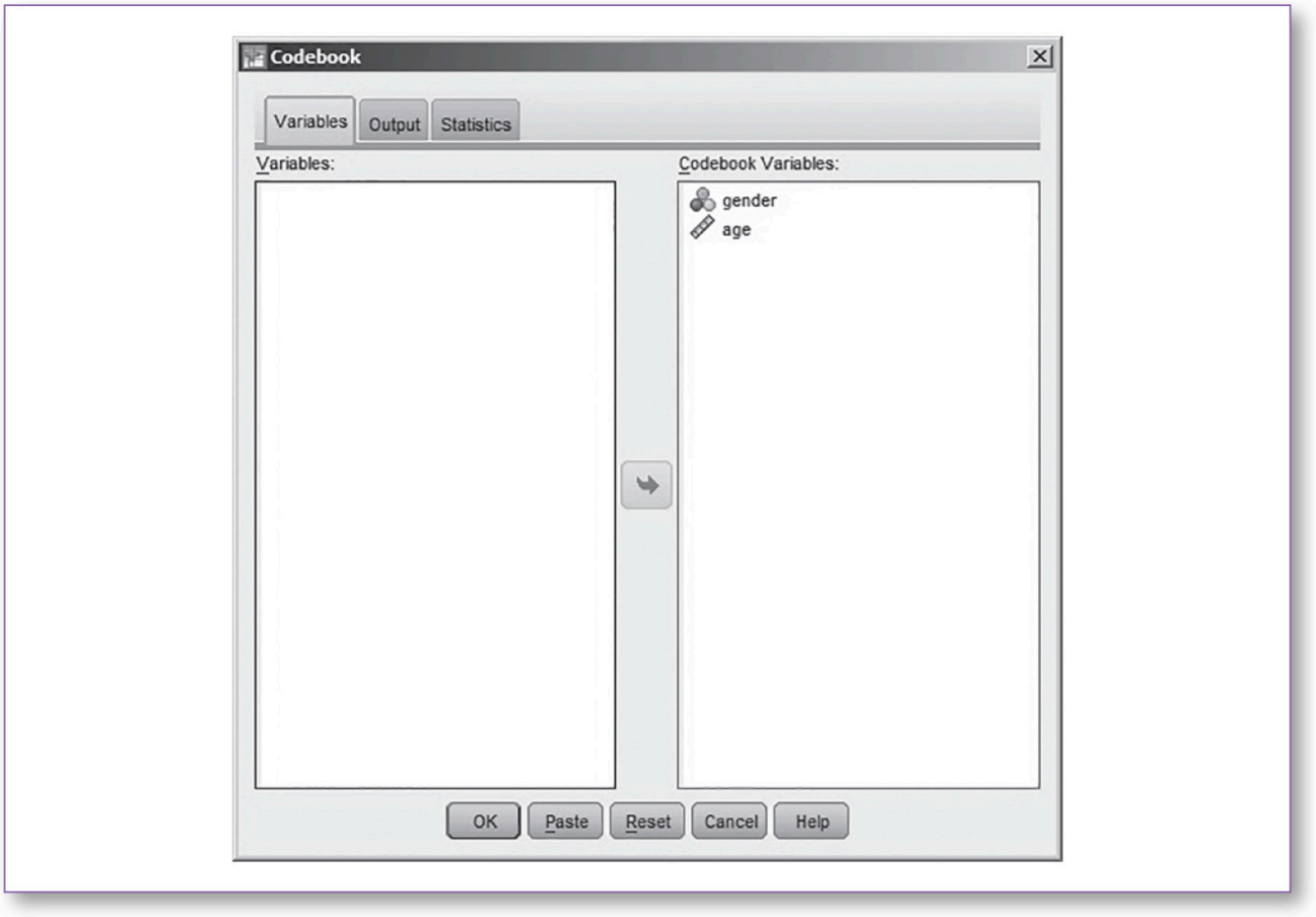


Table 3.2 Codebook Report Displaying the Variable Properties.

Table 3.2 Codebook Report Displaying the Variable Properties.

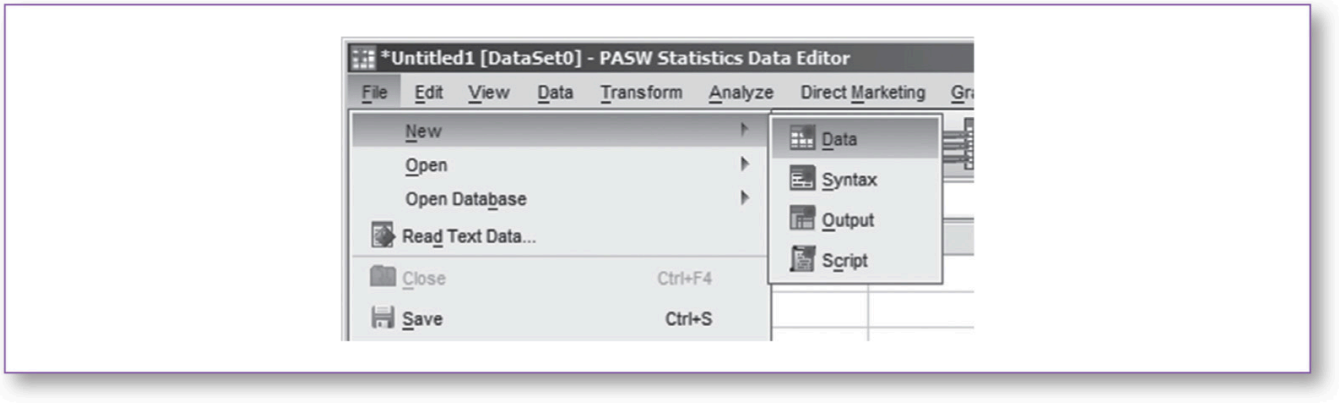
gender

		Value	Count	Percent
Standard Attributes	Position	1		
	Label	<none>		
	Type	Numeric		
	Format	F8		
	Measurement	Nominal		
	Role	Input		
Valid Values	1	Female	2	40.0%
	2	Male	3	60.0%

age

		Value
Standard Attributes	Position	2
	Label	<none>
	Type	Numeric
	Format	F8
	Measurement	Scale
	Role	Input
N	Valid	5
	Missing	0
Central Tendency and Dispersion	Mean	25.20
	Standard Deviation	4.382
	Percentile 25	24.00
	Percentile 50	25.00
	Percentile 75	27.00

Figure 3.11 Clearing the data, click on File, New, Data.



After each exercise, clear out the data; click on *File*, *New*, *Data*, as shown in Figure 3.11.

Exercise 3.1

Codebook

Variable: enrolled
Definition: Currently enrolled in school
Type: Categorical (1 = Yes, 2 = No)
Variable: units
Definition: Number of units the student is enrolled in
Type: Continuous
Variable: grade
Definition: Overall grade
Type: Continuous
Data:

	enrolled	units	grade
1	Yes	12	70

2	Yes	12	93
3	No	0	81
4	Yes	6	72
5	Yes	16	91

NOTE: Do not enter the numbers in the first shaded column (1, 2, 3, 4, and 5); this column corresponds to the leftmost column in SPSS, indicating the record (row) number.

Exercise 3.2

Codebook

Variable: id
Definition: ID number
Type: String
Variable: volunteer_hours
Definition: Volunteer hours worked per week
Type: Continuous
Data:

	id	volunteer_hours
1	QF732	2.00
2	AL331	1.50

3	JW105	3.00
4	RK122	.50
5	DD987	4.00

Exercise 3.3

Codebook

Variable: degree
Definition: Highest degree completed
Type: Categorical (1 = Associate's, 2 = Bachelor's, 3 = Master's, 4 = Doctorate)
Variable: pretest
Definition: Pretest score
Type: Continuous
Variable: posttest
Definition: Posttest score
Type: Continuous
Data:

	degree	pretest	posttest
1	Associate's	22	29
2	Master's	31	48
3	Bachelor's	28	38

4	Bachelor’s	25	34
5	Master’s	30	46

Exercise 3.4

Codebook

Variable: employ
Definition: Employment status
Type: Categorical (1 = Unemployed, 2 = Temporary, 3 = Part-time, 4 = Full-time)
Variable: work_hours
Definition: Average hours worked per week
Type: Continuous
Variable: sleep_hours
Question: Average sleep hours per day (0–24)
Type: Continuous
Data:

	employ	work_hours	sleep_hours
1	Unemployed	0	10.00
2	Temporary	16	9.00
3	Full-time	40	7.50
4	Full-time	45	8.00

5	Part-time	20	7.00

Exercise 3.5

Codebook

Variable: first_initial
Definition: First letter of first name
Type: String
Variable: last_name
Definition: Last name
Type: String
Variable: siblings
Definition: Number of brothers and sisters
Type: Continuous
Variable: adopted
Definition: Are you adopted?
Type: Categorical (1 = Yes, 2 = No)
Data:

	first_initial	last_name	siblings	adopted
1	J	Gower	0	No
2	D	Freeman	2	No
3	T	Rexx	3	No

4	P	Smith	2	Yes
5	V	Jones	1	No

Exercise 3.6

Codebook

Variable: patient_id
Definition: Patient ID number
Type: Continuous
Variable: age
Definition: Patient's age
Type: Continuous
Variable: temp
Definition: Body temperature (°F)
Type: Continuous
Variable: flu_shot
Definition: Has the patient had a flu shot this season?
Type: Categorical (1 = Yes; 2 = No, and I don't want one; 3 = Not yet, but I'd like one)
Variable: rx
Definition: Current medications
Type: String
Data:

	patient_id	age	temp	flu_shot	rx
1	2136578099	22	98.6	Yes	

2	8189873094	24	99.0	No, and I don’t want one	Multivitamin
3	2144538086	53	101.5	Not yet, but I want one	
4	8046628739	81	98.8	Yes	
5	5832986812	38	100.9	Yes	Xamine, Tutsocol

NOTE: Drug names are fictitious.

Exercise 3.7

Codebook

Variable: passport
Definition: Do you have a valid passport?
Type: Categorical (1 = Yes, 2 = No, 3 = Decline to answer)
Variable: fired
Definition: Have you ever been fired from a job?
Type: Categorical (1 = Yes, 2 = No, 3 = Decline to answer)
Variable: er
Definition: Have you ever been treated in an emergency room?
Type: Categorical (1 = Yes, 2 = No, 3 = Decline to answer)
Variable: dob
Definition: Date of birth
Type: Date
Data:

	passport	fired	er	dob

1	No	Decline to answer	No	01/23/1936
2	Yes	No	Yes	08/18/1928
3	Yes	No	No	03/01/1987
4	No	Yes	No	06/07/1974
5	No	No	Yes	11/30/2001

Exercise 3.8

Codebook

Variable: dogs
Definition: I like dogs.
Type: Categorical (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree)
Variable: cats
Definition: I like cats.
Type: Categorical (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree)
Variable: pets
Definition: How many pets do you currently have?
Type: Continuous
Data:

	dogs	cats	pets
1	Strongly agree	Disagree	1

2	Agree	Strongly agree	0
3	Strongly agree	Neutral	0
4	Strongly agree	Strongly agree	2
5	Neutral	Strongly disagree	3

Exercise 3.9

Codebook

Variable: blood_type
Definition: What is your blood type (respond “?” if you don’t know)?
Type: Categorical (1 = A−, 2 = A+, 3 = B−, 4 = B+, 5 = AB−, 6 = AB+, 7 = O−, 8 = O+, 9 = Don’t know)
Variable: gender
Definition: Gender
Type: Categorical (1 = Female, 2 = Male)
Variable: prior_donor
Definition: Have you ever donated blood before?
Type: Categorical (1 = Yes, 2 = No)
Data:

	blood_type	gender	prior_donor
1	B+	Female	Yes

2	Don’t know	Female	No
3	A−	Male	No
4	AB+	Male	No
5	O−	Male	Yes

Exercise 3.10

Codebook

Variable: entree_food
Definition: Entrée food
Type: Categorical (1 = Fish, 2 = Chicken, 3 = Beef, 4 = Vegetarian)
Variable: entree_quality
Definition: Entrée quality (1 = Poor . . . 10 = Excellent)
Type: Continuous
Variable: dessert_quality
Definition: Dessert quality (1 = Poor . . . 10 = Excellent)
Type: Continuous
Data:

	entree_food	entree_quality	dessert_quality
1	Fish	9	7
2	Fish	8	9

3	Fish	9	10
4	Beef	10	8
5	Fish	7	10

https://doi.org/10.4135/9781071878910