Chapter Two: Creating a Data File

This chapter explains how to set up a file with new data. After finishing this chapter, you should be able to create an IBM SPSS data file that will include the data and some labeling that gives more detail about the data. To illustrate this process, we will use a shortened version of the questionnaire used by the General Social Survey conducted by the National Opinion Research Center. For this example, our students wanted to see if their opinions on social issues were similar to those of the national sample.

The students knew they were not a representative sample, even of college students, but this questionnaire is an interesting way to learn how to create a new data file. They decided to use the following questions[1]:

- What is your age?

- Are you male or female?

- What is your religious preference?

- Generally speaking, in politics, do you consider yourself as conservative, liberal, or middle of the road?

- What kind of marriage do you think is the more satisfying way of life: one where the husband provides for the family and the wife takes care of the house and children or one where both the husband and wife have jobs and both take care of the house and children?

- Do you think it should be possible for a pregnant woman to obtain a <u>legal</u> abortion:

    - If there is a strong chance of a serious defect in the baby?

    - If she is married and does not want any more children?

    - If the woman's own health is seriously endangered by pregnancy?

    - If the family has a very low income and cannot afford any more children?

    - If she became pregnant as a result of rape?

    - If she is not married and does not want to marry the man?

    - If the woman wants it for any reason?

Basic Steps in Creating a Data File

It is best to start a data file with some careful planning.

1. First we will assign each respondent an identification number. This is not so we can identify individuals, but so we can keep track of each case when we go back to check the accuracy of the data entering. Each question is a variable in our data set. It needs a variable name that is simple but expresses something important about the data.

---

[1] A copy of this questionnaire is included as Appendix 2-A at the end of this chapter.

(IBM SPSS limits variable names to 64 characters or fewer. They may be numbers or letters but not spaces and very few special characters, so don't use any odd symbols.) *Age* and *sex* would be good variable names for the first two questions.[2] For the questions on abortion, we decided to use the first three characters of the variable names used by the General Social Survey. We used *mg* for the preferred type of marriage and called political orientation *conlib*. Each variable name can be given an extended variable label that gives more detail. (Extended variable labels can use spaces or special characters.) For example, *conlib* could have a variable label that said Conservative-Liberal.

2. After we have given each variable a name and label, we give each possible response to the question a code called a value label that is often the number corresponding to the order of the answers. (We could use another system, but this is the easiest because IBM SPSS works best with numeric codes to represent the data.) For example, *sex* could use 1 for male and 2 for female; *conlib* could use 1 for conservative, 2 for liberal, and 3 for middle of the road. Values would then be given value labels such as Male, Female, Conservative, Liberal, and Middle of the Road.

3. Sometimes respondents do not answer a question, give more than one answer, or do something else that makes their answers unusable. In our example, respondent #2 marked both yes and no on the last question, respondent #3 wrote in none on question 4, and respondent #13 didn't answer the marriage question. We assign these missing value codes so they don't distort the analysis. Often 9 is used to indicate missing data or 99 if it is a two-digit value.

Everything must be planned carefully before entering the data into IBM SPSS. It is useful to put the data in a matrix like Table 2.1 before entering it into the IBM SPSS Data Editor. For this exercise, we will use only the first four questions and five respondents. (The complete matrix is Appendix 2.B at the end of this chapter.)

Table 2.1 Matrix for Data-entry Exercise

id	age	Sex	rel	conlib
01	20	1	4	2
02	24	2	5	2
03	21	2	2	9
04	24	2	5	3
05	26	2	4	2

---

[2] For this exercise, we used lower-case italics for the variable names.

Getting Started in IBM SPSS

To create the data file in IBM SPSS, open IBM SPSS (probably by clicking on the IBM SPSS icon on the desktop). (See Figure 2-1.)

This opens a matrix similar to a spreadsheet such as Excel or the matrix we just worked on. The rows will be the cases (the respondents) and the columns will be the variables (answers to the questions). So, the upper-left cell will contain an identification number for the first case and the cells to the right will be data about that case. The IBM SPSS Data Editor has tabs in the lower-left corner that let you work with your data in two ways. Variable View is used to set up the data—names, variable labels, value labels, etc. The other tab, Data View, is used to actually enter the data. IBM SPSS probably opened in the Data View mode, if not, click the Data View tab at the bottom left of the IBM SPSS screen now. (See Figure 2-2.)

Entering Variable and Value Names and Labels:

In Data View, we will use the first column for the respondents' ID numbers, so type **001** into the first cell and press Enter and 1.00 will appear.   (See Figure 2-3.) [3]

We will use the Variable View tab to assign variable names and longer variable labels as well as value names and labels that will make it easier to use the data for tables and charts. Click Variable View now and click the VAR0001 in the top left column. Type in id. (Press Enter and VAR0001 changes to our variable name, id.) Go back to Data View and notice that the first column is now titled id. (See Figure 2-4.)
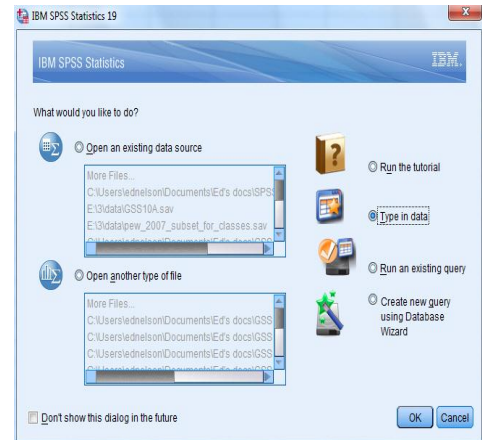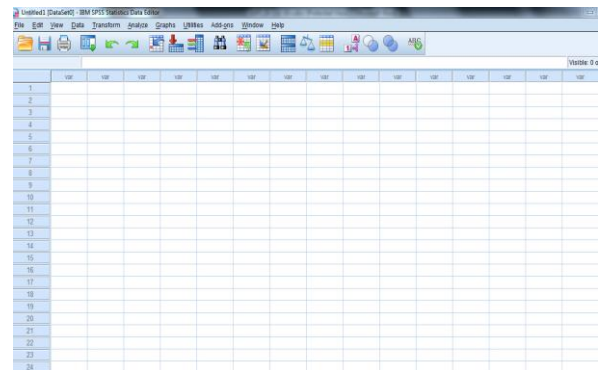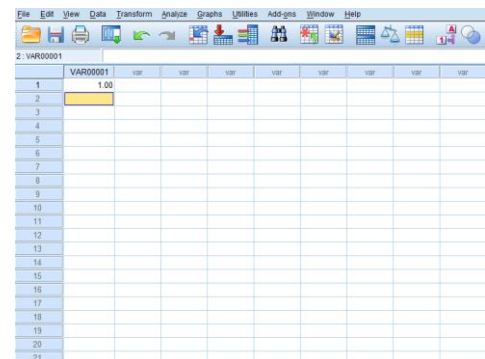

Figure 2-1

Figure 2-2

Figure 2-3

---

[3] It is wise to save your computer work early and often. You might want to save this file and call it something like **Data Entry Exercise 1**. Notice that IBM SPSS saves it as a .sav file. This means it contains the data in the format for IBM SPSS analysis.

The second variable will be the student's age, so change back with the Variable View and type **age** under name in the second row. Remember to press Enter after you type in each entry.

IBM SPSS makes some assumptions about data that might not be appropriate. In the fourth column, notice that it plans to use two decimal points even when the values for age are integers. To avoid these inappropriate decimal points, select the decimals column by clicking the heading and use the blue box to go down to 0. (Remember to do this whenever a numeral doesn't really refer to a numerical value.)
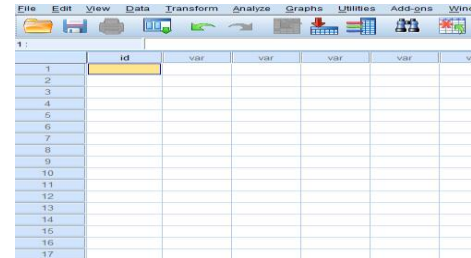

Figure 2-4

Since the short variable name usually doesn't give enough information about the variable, we want a longer or clearer variable label for our analysis. This one would be simple. To add a variable label to *age*, just tab over to the label column and type in **Age**. (See Figure 2-5.) Although, it may not seem necessary to have a variable label for age, for most variables a longer variable label is very useful.
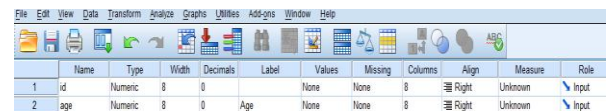

Figure 2-5

Sometimes respondents don't answer a question, give two answers, or do something else so the data can't be used in the analysis. To have accurate results, missing or invalid data need to be indicated. Still in Variable View, tab over to Missing and click the blue box. This dialog box lets you specify up to three missing values. For our data, click Discrete and type **99** in the first text box. Leave the other boxes empty. Then Click OK. Now if someone doesn't answer a question, it will be marked as missing. (See Figure 2-6.)
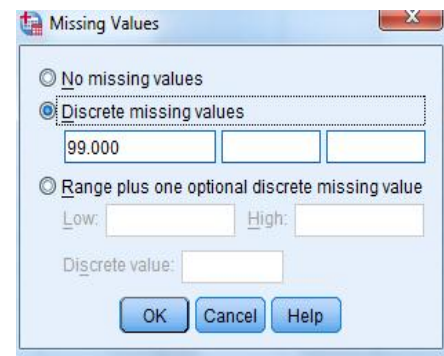

Figure 2-6

The third variable will be the sex of the respondent, so type **sex** in the third row under Name and Sex as the variable label. Since we're going to use the code 1 for males and 2 for females, we're going to need value names in words for each category. Tab over to the cell under Values and click the little blue box to get the value labels menu. Type a **1** in the Value box and then **Male** in the Value Label box. Click Add and it shows that Value 1 will be Male. Type a **2** in the Value box, and type **Female** in the Value
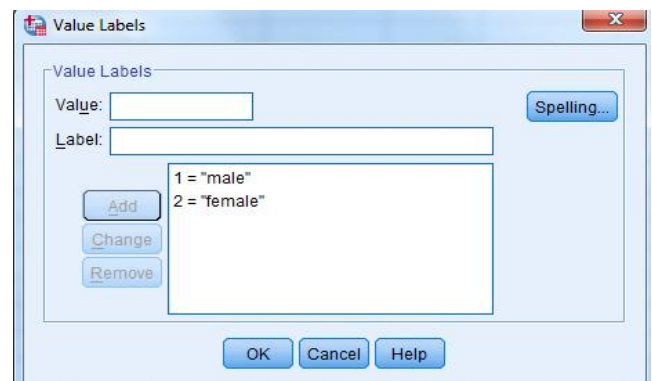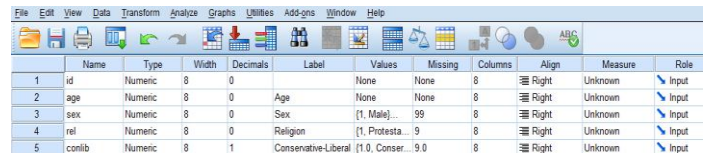

Figure 2-7

label space. Click Add and then click OK to save these. Now, IBM SPSS knows that 1 and 2 in Sex are really male and female respectively. (See Figure 2-7.)

For this exercise, we are also using religion and conservative-liberal as variables. Add those variables in rows 4 and 5. Give each variable a name and label—*rel* gets Religion and *conlib* gets something like Conservative-Liberal as variable name and label. Then add value names and labels.

Notice that *rel* has five possibilities—Protestant, Catholic, Jewish, other, and no religion. Go ahead and work out the value names and value labels. Make arrangements for missing values just as you did

Figure 2-8

before. (You can refer to Appendix 2-A, Codebook for Student Questionnaire at the end of this chapter.) Remember to type variable labels, value names, and value labels exactly the way you would want them in a table when you do the analysis—often this is with the first letter of each important word capitalized. (Your Variable View might look like Figure 2-8.)

Entering the Data:

Enter the codes for each variable using Data View[4]. Then check the accuracy of your data entry by scanning down each column looking for codes that would be impossible. For example, sex can have only three possibilities since male is 1, female is 2, and missing information is 9, so a 5 or 6 would be a mistake. Then check everything carefully. The best check is to have one person read the codes while another checks the entries on Data View.

---

[4] Some people, especially those who are used to working with spreadsheets, like to enter all the data in Data View before they set up the variable names, etc. In this example, we set up the variable names, etc., before we enter any data. (You'll have to figure out what works best for you.) You can also enter data from a spreadsheet like Excel.

Student Survey Questionnaire

(1) What is your age? _____

(2) Are you ____ male or ___ female?

(3) What is your religious preference?

___ Protestant ___Catholic ___ Jewish ___ Some other religion ___No religion

(4) Generally speaking, in politics, do you consider yourself as

___conservative ___ liberal __ or middle of the road

(5) What kind of marriage do you think is the more satisfying way of life?

___ One where the husband provides for the family and the wife takes care of the house and children

___ One where both the husband and wife have jobs and both take care of the house and children

Do you think it should be possible for a pregnant woman to obtain a <u>legal</u> abortion?

(6) If there is a strong chance of serious defect in the baby?
__Yes __ No ___Don't Know

(7) If she is married and does not want any more children?
__Yes __ No ___Don't Know

(8) If the woman's own health is seriously endangered by pregnancy?

__Yes __ No ___Don't Know

(9) If the family has a very low income and cannot afford any more children?

__Yes __ No ___Don't Know

(10) If she became pregnant as a result of rape?

__Yes __ No ___Don't Know

(11)   If she is not married and does not want to marry the man?
__Yes __No __ Don't Know

(12)   If the woman wants it for any reason
__Yes __ No ___Don't Know

Appendix 2A: Codebook for Student Questionnaire

Missing Values	9 or 99
Age	Age at last birthday
Sex	1 = male, 2 = female
Religious Preference	1 = Protestant, 2 = Catholic, 3 = Jewish, 4 = Other, 5 = None
Political Orientation	1 = Conservative, 2 = Liberal, 3 = Middle of the road
Preferred Marriage	1 = Traditional, 2 = Shared
Abortion if Birth Defect	1= Yes, 2 = No, 3 = Don't Know
Abortion if No More Children	1= Yes, 2 = No, 3 = Don't Know
Abortion if Health Risk	1= Yes, 2 = No, 3 = Don't Know
Abortion if Poor	1= Yes, 2 = No, 3 = Don't Know
Abortion if Rape	1= Yes, 2 = No, 3 = Don't Know
Abortion if Not Married	1= Yes, 2 = No, 3 = Don't Know
Abortion For Any Reason	1= Yes, 2 = No, 3 = Don't Know

Appendix 2B: Planning Matrix for Data-entry Exercise

	age	sex	rel	c-l	mg	abd	abn	abh	abp	abr	abs	aba
01	20	1	4	2	2	2	2	1	3	1	2	2
02	24	2	5	2	2	1	1	1	1	1	1	9
03	21	2	2	9	2	2	2	2	2	2	2	2
04	24	2	5	3	2	1	1	1	1	1	1	1
05	26	2	4	2	2	1	1	1	1	1	1	1
06	28	2	2	2	2	2	2	1	2	1	2	2
07	23	1	1	2	2	1	2	1	1	1	2	2
08	22	2	4	3	1	1	1	1	1	1	1	1
09	22	1	5	2	2	1	1	1	1	1	1	1
10	22	2	4	4	2	1	1	1	1	1	1	1
11	23	1	2	2	1	2	2	1	2	1	2	3
12	24	2	2	3	2	1	1	1	1	1	1	2
13	51	2	1	2	9	1	1	1	1	1	1	1
14	22	2	2	3	2	1	1	1	1	1	1	1
15	21	2	4	3	2	1	1	1	1	1	1	1
16	37	1	1	3	2	1	2	1	2	1	2	2
17	22	2	4	2	2	1	1	1	1	1	2	2
18	22	2	3	3	2	1	2	1	2	1	2	2
19	22	2	4	3	2	3	2	1	2	1	1	1
20	30	2	5	2	2	1	1	1	1	1	1	1
21	25	2	5	2	2	1	1	1	1	1	1	1
22	23	1	2	2	2	1	1	1	1	1	1	1
23	21	1	1	2	1	1	1	2	1	2	1	1

Chapter Two Exercises

Exercise 2-1. Clients of Friendly Visitor Service

At California State University, Fresno, the Friendly Visitors Service hires college students to do in-home care for elderly people so they can remain independent and stay in their homes as long as possible. The students do cleaning, yard work, shopping, etc. The staff begins by interviewing clients in their homes and assessing their need for services. The following information is used to match the seniors with the students who want employment:

- Age: Age at Last Birthday
- Sex: Male or Female
- Lives alone: Yes or No
- Low income: Yes = Eligible for Supplemental Security Income (SSI)
- Need for assistance with the activities of daily living (ADL): Bathing, dressing, toileting, transferring in/out of bed, eating
- Total number of ADLs needing help
- Need for assistance with the instrumental activities of daily living (IADL): Using telephone, shopping, preparing food, light housework, heavy housework, finances
- Total Number IADLs needing help

To keep track of the needs of potential clients, the program could create a data file and use it in IBM SPSS. (Data from one month's new applications are provided below. For this example, we'll just use the count of the number of activities for which the seniors need help, but note that they could include the yes/no responses for each of the activities of daily living.)

Sample Data Set: Friendly Visitor Service Clients

id	age	sex	alone	income	#adl	#iadl
001	74	M	N	N	0	4
002	66	M	N	N	4	6
003	81	M	N	N	2	5
004	76	F	N	N	0	4
005	74	M	N	N	1	5
006	69	F	N	Y	0	4
007	79	F	Y	N	0	4
008	80	M	N	Y	3	6
009	89	M	N	N	3	5
010	60	F	Y	N	2	6
011	88	F	Y	N	0	3
012	82	F	Y	N	2	4
013	79	F	Y	N	1	4
014	77	M	N	N	3	6
015	62	M	Y	N	1	4
016	83	M	N	N	4	6
017	80	F	Y	N	0	2
018	85	F	N	N	1	4
019	66	F	Y	N	1	3
020	84	M	N	N	4	6
021	74	F	N	N	4	4
022	74	M	N	N	0	2
023	74	F	Y	N	0	5
024	92	M	N	N	3	6
025	66	F	N	N	2	6

Exercise 2-2. Age at Death from Newspaper Obituaries

An interesting source of data for student practice with data analysis using IBM SPSS is the death notices in local newspapers. Although big city newspapers publish obituaries only for the rich and famous, many local newspapers provide information on almost everyone who dies in the community. (For example, see *The Fresno Bee* at www.fresnobee.com/obituaries, which publishes information provided by funeral homes on most deaths in the community.) Click on the date to get the alphabetized list of people for that day. From these death notices, you could set up a data file with the age and sex of each person who died at a particular time (for example, the first month in the term). The age or birthday is usually given, and you can infer sex from names or pronouns. (The longer, more-detailed obituaries provided by some families would not be a suitable sample for a statistical analysis.) We could use this with IBM SPSS for an analysis of age and sex at death, for example, obtaining frequency and percent distributions, various charts, and descriptive statistics in Chapter 4; cross tabulations in Chapter 5, and/or comparison of means in Chapter 6.