Chapter 5

Learning SPSS: Data and EDA

An introduction to SPSS with emphasis on EDA.

SPSS (now called PASW Statistics, but still referred to in this document as SPSS) is a perfectly adequate tool for entering data, creating new variables, performing EDA, and performing formal statistical analyses. I don't have any special endorsement for SPSS, other than the fact that its market dominance in the social sciences means that there is a good chance that it will be available to you wherever you work or study in the future. As of 2009, the current version is 17.0, and class datasets stored in native SPSS format in version 17.0 may not be usable with older versions of SPSS. (Some screen shots shown here are not updated from previous versions, but all changed procedures have been updated.)

For very large datasets, SAS tends to be the best program. For creating custom graphs and analyses R, which is free, or the commercial version, S-Plus, are best, but R is not menu-driven. The one program I strongly advise against is Excel (or any other spreadsheet). These programs have quite limited statistical facilities, discourage structured storage of data, and have no facility for documenting your work. This latter deficit is critical! For any serious analysis you must have a complete record of how you created new variables and produced all of your graphical and statistical output.

It is *very* common that you will find some error in your data at some point. So it is highly likely that you will need to repeat all of your analyses, and that is painful without exact records, but easy or automatic with most good software. Also, because it takes a long time from analysis to publishing, you will need these

records to remind yourself of exactly which steps you performed.

As hinted above, the basic steps you will take with most experimental data are:

1. Enter the data into SPSS, or load it into SPSS after entering it into another program.
2. Create new variables from old variables, if needed.
3. Perform exploratory data analyses.
4. Perform confirmatory analyses (formal statistical procedures).
5. Perform model checking and model comparisons.
6. Go back to step 4 (or even 2), if step 5 indicates any problems.
7. Create additional graphs to communicate results.

Most people will find this chapter easier to read when SPSS is running in front of them. There is a lot of detail on getting started and basic data management. This is followed by a brief compilation of instructions for EDA. The details of performing other statistical analyses are at the end of the appropriate chapters throughout this book.

Even if you are someone who is good at jumping in to a computer program without reading the instructions, I urge you to read this chapter because otherwise you are likely to miss some of the important guiding principles of SPSS.

Additional SPSS resources may be found at
http://www.stat.cmu.edu/∼hseltman/SPSSTips.html.

5.1 Overview of SPSS

SPSS is a multipurpose data storage, graphical, and statistical system. At (almost) all times there are two window types available, the Data Editor window(s) which each hold a single data "spreadsheet", and the Viewer window from which analyses are carried out and results are viewed.

The Data Editor has two views, selected by tabs at the bottom of the window. The Data View is a spreadsheet which holds the data in a rectangular format with

cases as rows and variables as columns. Data can be directly entered or imported from another program using menu commands. (Cut-and-paste is possible, but not advised.) Errors in data entry can also be directly corrected here.

You can also use menu commands in the Data View to create new variables, such as the log of an existing variable or the ratio of two variables.

The Variable View tab of the Data Editor is used to customize the information about each variable and the way it is displayed, such as the number of decimal places for numeric variables, and the labels for categorical variables coded as numbers.

The Viewer window shows the results of EDA, including graph production, formal statistical analyses, and model checking. Most data analyses can be carried out using the menu system (starting in either window), but some uncommon analyses and some options for common analyses are only accessible through "Syntax" (native SPSS commands). Often a special option is accessed by using the Paste button found in most main dialog boxes, and then typing in a small addition. (More details on these variations is given under the specific analyses that require them.)

All throughout SPSS, each time you carry out a task through a menu, the underlying non-menu syntax of that command is stored by SPSS, and can be examined, modified and saved for documentation or reuse. In many situations, there is a "Paste" button which takes you to a "syntax window" where you can see the underlying commands that would have been executed had you pressed OK.

SPSS also has a complete help system and an advanced scripting system.

You can save data, syntax, and graphical and statistical output separately, in various formats whenever you wish. (Generally anything created in an earlier program version is readable by later versions, but not vice versa.) Data is normally saved in a special SPSS format which few other programs can understand, but universal formats like "comma separated values" are also available for data interchange. You will be warned if you try to quit without saving changes to your data, or if if you forget to save the output from data analyses.

As usual with large, complex programs, the huge number of menu items available can be overwhelming. For most users, you will only need to learn the basics of interaction with the system and a small subset of the menu options.

Some commonly used menu items can be quickly accessed from a toolbar, and learning these will make you more efficient in your use of SPSS.

SPSS has a few quirks; most notably there are several places where you can make selections, and then are supposed to click Change before clicking OK. If you forget to click Change your changes are often silently forgotten. Another quirk that is well worth remembering is this: *SPSS uses the term Factor to refer to any categorical explanatory variable.* One good "quirk" is the Dialog Recall toolbar button. It is a quick way to re-access previous data analysis dialogs instead of going through the menu system again.

5.2 Starting SPSS

Note: SPSS runs on Windows and Mac operating systems, but the focus of these notes is Windows. If you are unfamiliar with Windows, the link
Top 10 tips for Mac users getting started with Windows may help.

Assuming that SPSS is already installed on your computer system, just choose it from the Windows Start menu or double click its icon to begin. The first screen you will see is shown in figure 5.1 and gives several choices including a tutorial and three choices that we will mainly use: "Type in data", "Open an existing data source", and "Open another type of file". "Type in data" is useful for analyzing small data sets not available in electronic form. "Open an existing data source" is used for opening data files created in SPSS. "Open another type of file" is used for importing data stored in files not created by SPSS. After making your choice, click OK. Clicking Cancel instead of OK is the same as choosing "Type in data".

Use Exit from the File menu whenever you are ready to quit SPSS.

5.3 Typing in data

To enter your data directly into SPSS, choose "Type in data" from the opening screen, or, if you are not at the opening screen, choose New then Data from the File menu.

The window titled "Untitled SPSS Data Editor" is the Data Editor window which is used to enter, view and modify data. You can also start statistical analyses from this window. Note the tabs at the bottom of the window labeled "Data View" and "Variable View". In Data View (5.2), you can view, enter, and edit data for all of your cases, while in Variable View (5.3), you can view, enter, and edit

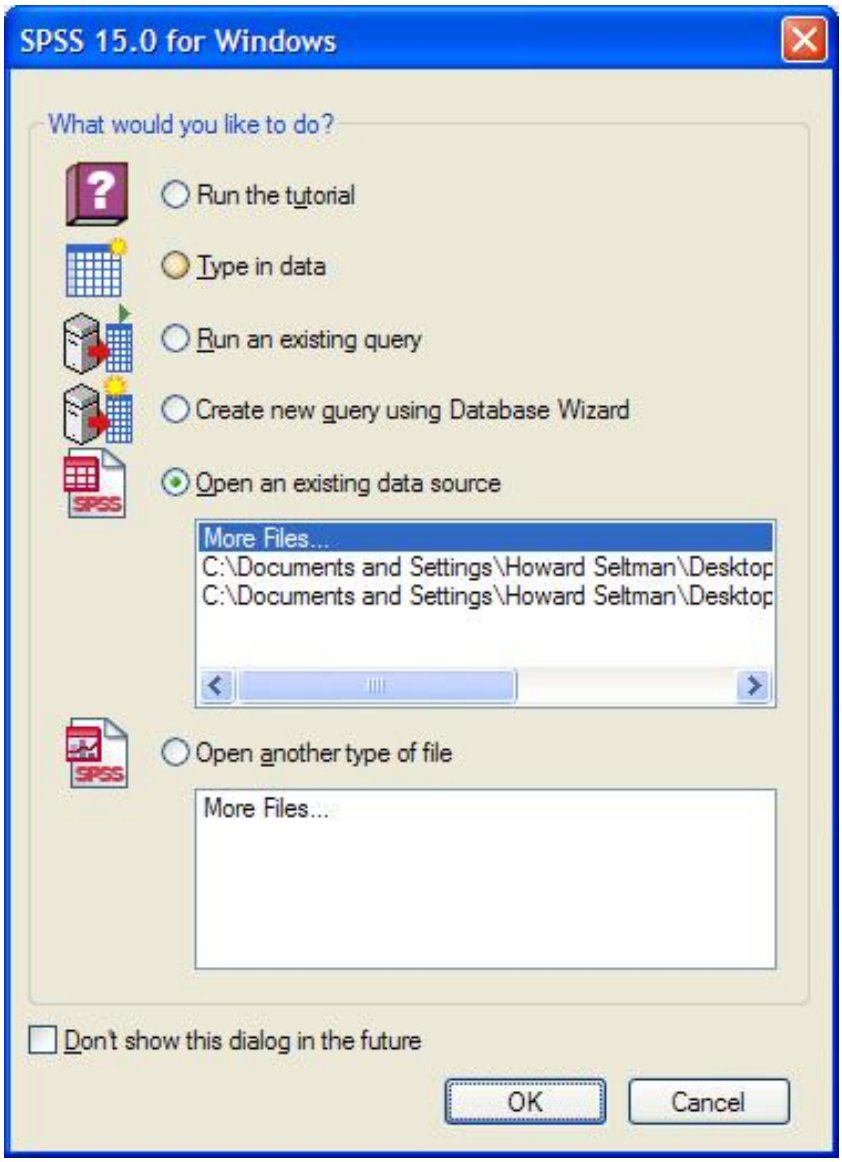

Figure 5.1: SPSS intro screen.

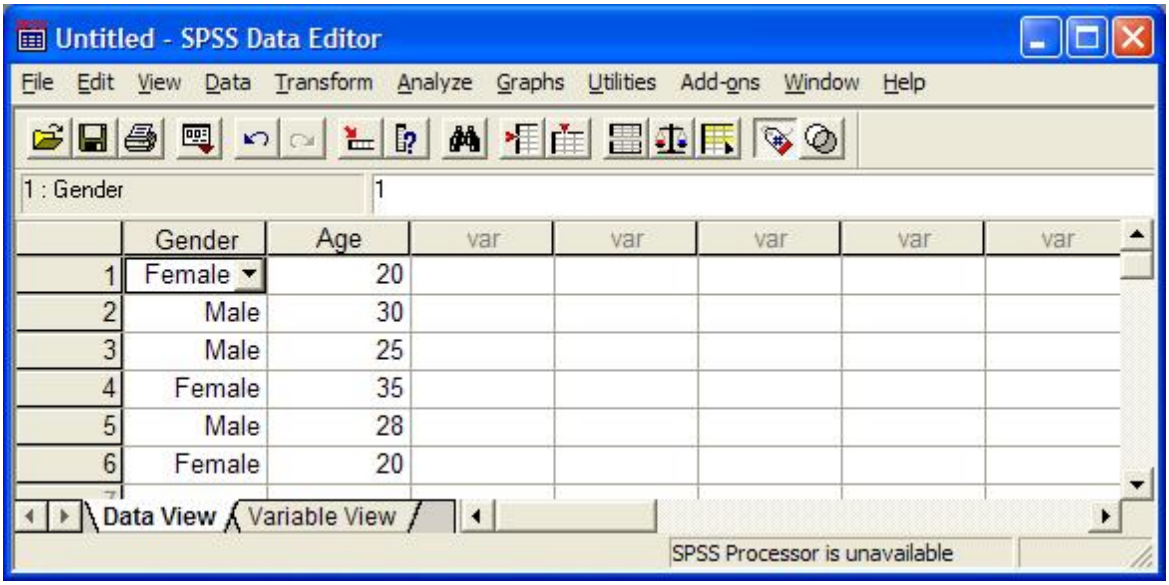

Figure 5.2: Data Editor window: Data View.

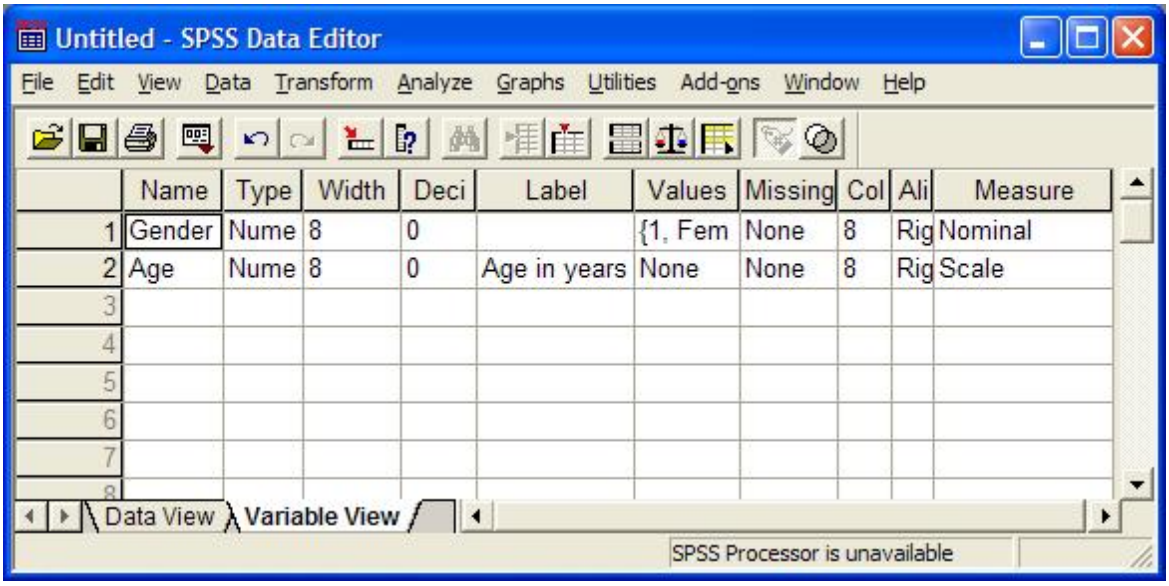

Figure 5.3: Data Editor window: Variable View.

information about the variables themselves (see below). Also note the menu and toolbar at the top of the window. You will use these to carry out various tasks related to data entry and analysis. There are many more choices than needed by a typical user, so don't get overwhelmed! You can hover the mouse pointer over any toolbar button to get a pop-up message naming its function. This chapter will mention useful toolbar items as we go along. (Note: Toolbar items that are inappropriate for the current context are grayed out.)

Before manually entering data, you should tell SPSS about the individual variables, which means that you should think about variable types and coding before entering the data. Remember that the two data types are categorical and quantitative and their respective subtypes are nominal and ordinal, and discrete and continuous. These data type correspond to the Measure column in the Variable View tab. SPSS does not distinguish between discrete and continuous, so it calls all quantitative variables "scale". Ordinal and nominal variables are the other options for Measure. In many parts of SPSS, you will see a visual reminder of the Measure of your variables in the form of icons. A small diagonal yellow rule indicates a "scale" variable (with a superimposed calendar or clock if the data hold dates or times). A small three level bar graph with increasing bar heights indicates an "ordinal" variable. Three colored balls with one on top and two below indicates nominal data (with a superimposed "a" if the data are stored as "strings" instead of numbers).

Somewhat confusingly SPSS Variable View has a column called Type which is the "computer science" type rather than the "statistics" data type. The choices are basically numeric, date and string with various numeric formats. This course does not cover time series, so we won't use the "date" Type. Probably the only use for the "string" Type is for alphanumeric subject identifiers (which should be assigned "nominal" Measure). All standard variables should be entered as numbers (quantitative variables) or numeric codes (categorical variables). Then, for categorical variables, we always want to use the Values column to assign meaningful labels to the numeric codes.

Note that, in general, to set or change something in the Data Editor, you first click in the cell whose row and column correspond to what you want to change, then type the new information. To modify, rather than fully re-type an entry, press the key labeled "F2".

When entering a variable name, note that periods and underscores are allowed in variable names, but spaces and most other punctuation marks are not. The

variable name must start with a letter, may contain digits, and must not end with a period. Variable names can be at most 64 characters long, are *not* case sensitive, and must be unique. The case that you enter is preserved, so it may be useful to mix case, e.g., hotDogsPerHour to improve readability.

In either View of the Data Editor, you can neaten your work by dragging the vertical bar between columns to adjust column widths.

After entering the variable name, change whichever other column(s) need to be changed in the Variable View. For many variables this includes entering a Label, which is a human-readable alternate name for each variable. It may be up to 255 characters long with no restrictions on what you type. The labels replace the variable names on much of the output, but the names are still used for specifying variables for analyses.

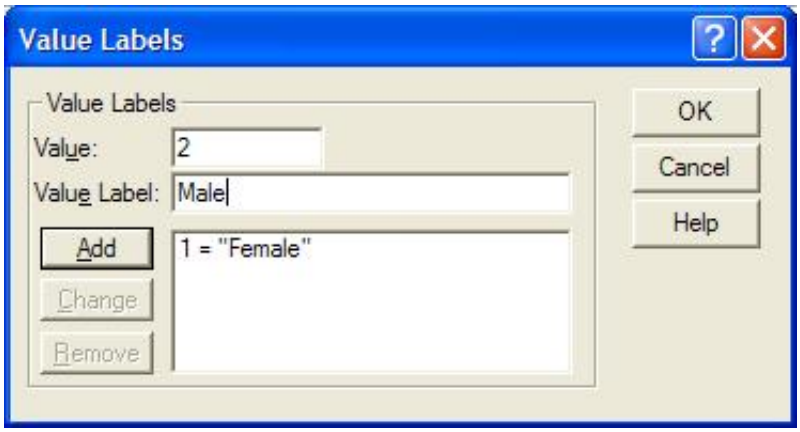

Figure 5.4: Values dialog box.

For categorical variables, you will almost always enter the data as numeric codes (Type "numeric"), and then enter Labels for each code. The Value Labels dialog box (5.4) is typical of many dialog boxes in SPSS. To enter Values for a variable, click in the box at the intersection of the variable's row and the Value column in the Variable View. Then click on the "..." icon that appears. This will open the "Value Labels" dialog box, into which you enter the words or phrases that label each level of your categorical variable. Value labels can contain anything you like up to 255 characters long. Enter a level code number in the Value box, press Tab, then enter the text for that level in the Value Label box. Finally you *must* click the Add button for your entry to be registered. Repeat the process as many times as needed to code all of the levels of the variable. When you are finished, verify

that all of the information in the large unlabeled box is correct, then click OK to complete the process. At any time while in the Value Label box (initially or in the future), you can add more labels; delete old labels by clicking on the variable in the large box, then clicking the Delete button; or change level values or labels by selecting the variable in the large box, making the change, then clicking the Change button. Version 16 has a spell check button, too.

If your data has missing values, you should use the Missing column of the Variable View to let SPSS know the missing value code(s) for each variable.

The only other commonly used column in Variable View is the Measure column mentioned above. SPSS uses the information in the column sporadically. Sometimes, but certainly not always, you will not be able carry out the analysis you want if you enter the Measure incorrectly (or forget to set it). In addition, setting the Measure assures that you appropriately think about the type of variable you are entering, so it is a really, really good idea to always set it.

Once you have entered all of the variable information in Variable View, you will switch to Data View to enter the actual data. At it's simplest, you can just click on a cell and type the information, possibly using the "F2" key to edit previously entered information. But there are several ways to make data entry easier and more accurate. The tab key moves you through your data case by case, covering all of the variables of one case before moving on to the next. Leave a cell blank (or delete its contents) to indicate "missing data"; missing data are displayed with a dot in the spreadsheet (but don't type a dot).

The Value Labels setting, accessed either through its toolbar button (which looks like a gift tag) or through the View menu, controls both whether columns with Value Labels display the value or the label, and the behavior of those columns during data entry. If Value Labels is turned on, a "..." button appears when you enter a cell in the Data View spreadsheet that has Value Labels. You can click the button to select labels for entry from a drop down box. Also, when Value Labels is on, you can enter data either as the code or by typing out the label. (In any case the code is what is stored.)

You should use Save (or Save as) from the File menu to save your data after every data entry session and after any edits to your data. Note that in the "Save Data As" dialog box (5.5) you should be careful that the "Save in:" box is set to save your data in the location you want (so that you can find it later). Enter a file name and click "Save" to save your data for future use. Under "Save as type:" the default is "SPSS" with a ".sav" extension. This is a special format that

can be read quickly by SPSS, but not at all by most other programs. For data exchange between programs, several other export formats are allowed, with Excel with "Comma separated values" being the most useful.

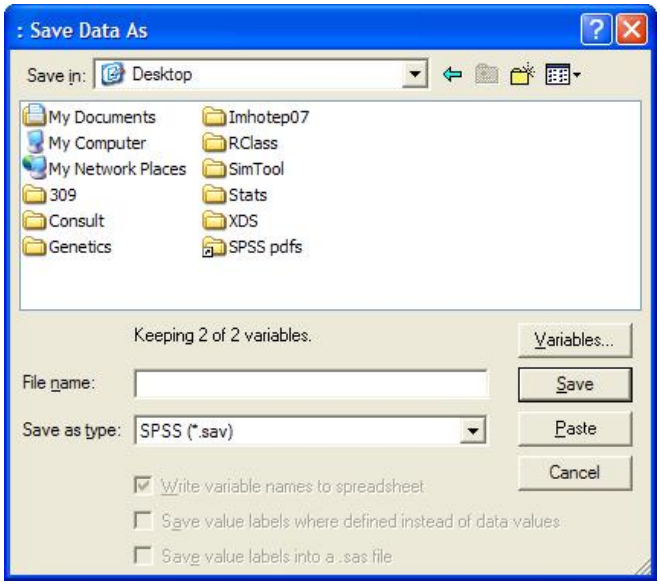

Figure 5.5: Save Data As dialog box.

5.4 Loading data

To load in data when you first start SPSS, your can select a file in one of the two lower boxes of the "Intro Screen". At any other time you can load data from the File menu by selecting Open, then Data. This opens the "Open File" dialog box (5.6).

It's a good idea to save any changes to any open data set before opening a new file. In the Open File dialog box, you need to find the file by making appropriate choices for "Look in:" and "Files of type:". If your file has a ".txt" extension and you are looking for files of type ".dat", you will not be able to find your file. As a last resort, try looking for files of type "all files(*.*)". Click Open after finding your file.

If your file is a native SPSS ".sav" file, it will open immediately. If it is of another type, you will have to go through some import dialogs. For example, if

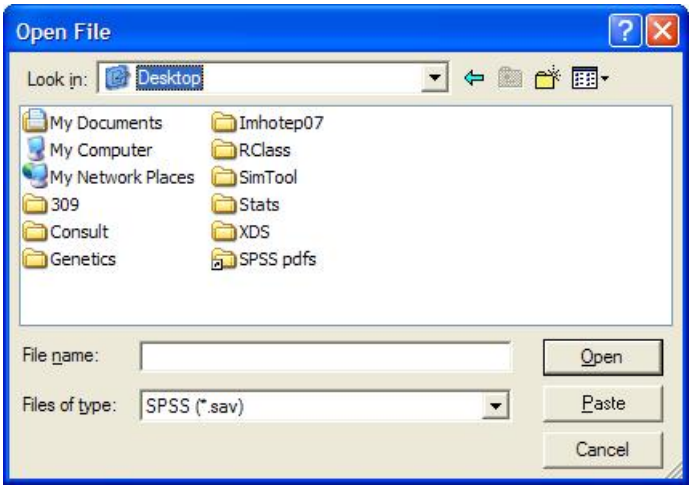

Figure 5.6: Open File dialog box.

you open an Excel file (.xls), you will see the "Opening Excel Data Source" dialog box (5.7). Here you use a check box to tell SPSS whether or not your data has variable names in the first row. If your Excel workbook has multiple worksheets you must select the one you want to work with. Then, optionally enter a Range of rows and columns if your data does not occupy the entire range of used cells in the worksheet. Finish by clicking OK.

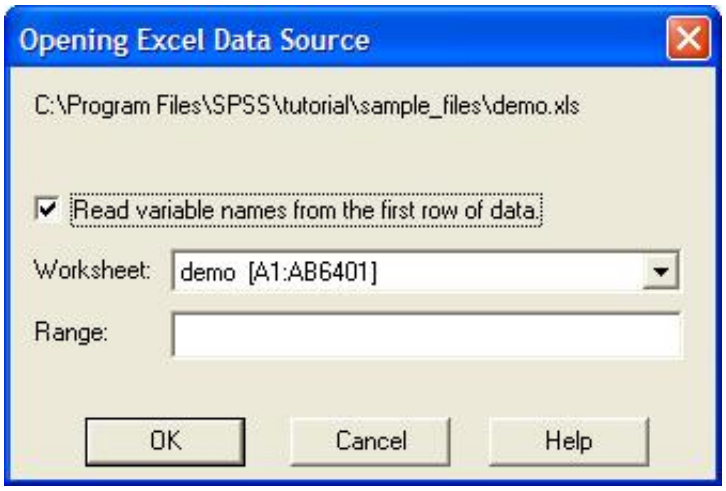

Figure 5.7: Open Excel Data Source dialog box.

The other useful type of data import is one of the simple forms of human-readable text such as space or tab delimited text (usually .dat or .txt) or comma separated values (.csv). If you open one of these files, the "Text Import Wizard" dialog box will open. The rest of this section describes the use of the text import wizard.

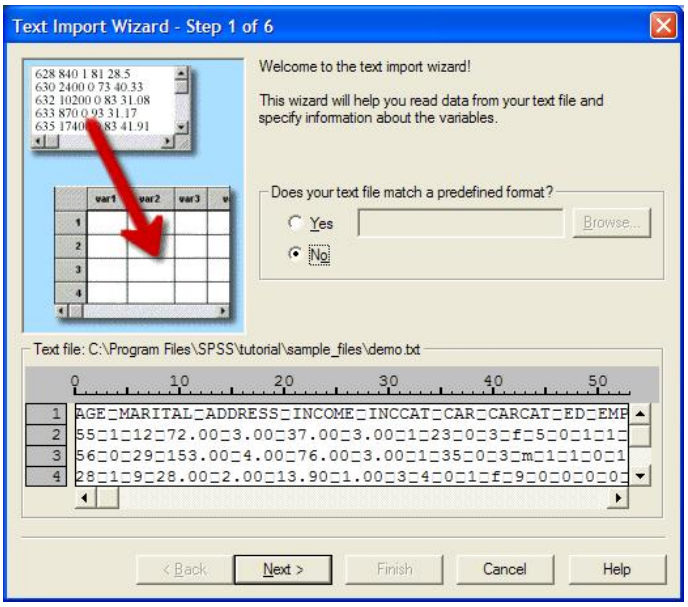

Figure 5.8: Text Import Wizard - Step 1 of 6.

In "Step 1 of 6" (5.8) you will see a question about predefined formats which we will skip (as being beyond the scope of this course), and below you will see some form of the first four lines of your file (and you can scroll down or across to see the whole file). (If you see strange characters, such as open squares, your file probably has non-printable characters such as tab character in it.) Click Next to continue.

In "Step 2 of 6" (5.9) you will see two *very important* questions that you must answer accurately. The first is whether your file is arranged so that each data column always starts in exactly the same column for every line of data (called "Fixed width") or whether there are so-called delimiters between the variable columns (also called "fields"). Delimiters are usually either commas, tab characters or one or more spaces, but other delimiters occasionally are seen. The second question is "Are variable names include at the top of the file?" Answer "no" if the first

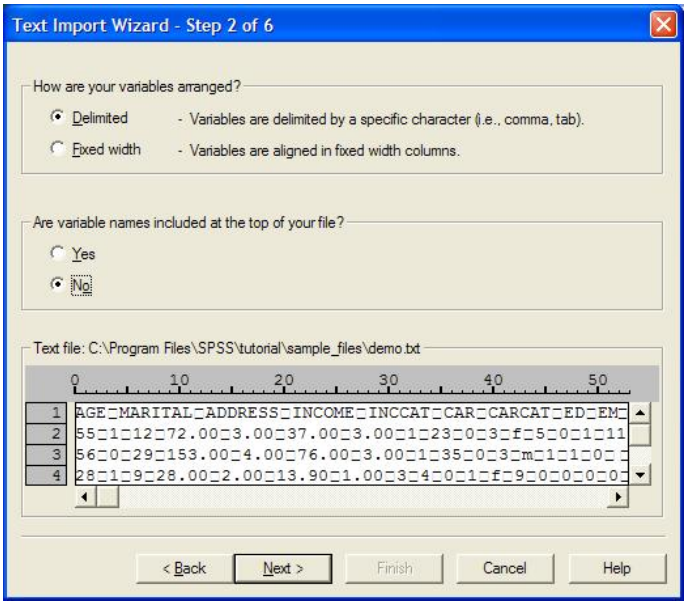

Figure 5.9: Text Import Wizard - Step 2 of 6.

line of the file is data, and "yes" if the first line is made of column headers. After answering these questions, click Next to continue.

In "Step 3 of 6" (5.10) your first task is to input the line number of the file that has the first real data (as opposed to header lines or blank lines). Usually this is line 2 if there is a header line and line 1 otherwise. Next is "How are your cases represented?" Usually the default situation of "Each line represents a case" is true. Under "How many cases do you want to import?" you will usually use the default of "All of the cases", but occasionally, for very large data sets, you may want to play around with only a subset of the data at first.

In "Step 4 of 6" (5.11) you must answer the questions in such a way as to make the "Data preview" correctly represent your data. Often the defaults are OK, but not always. Your main task is to set the delimiters between the data fields. Usually you will make a single choice among "Tab", "Space", "Comma", and "Semicolon". You may also need to specify what sets off text, e.g. there may be quoted multi-word phrases in a space separated file.

If your file has fixed width format instead of delimiters, "Step 4 of 6" has an alternate format (5.12). Here you set the divisions between data columns.

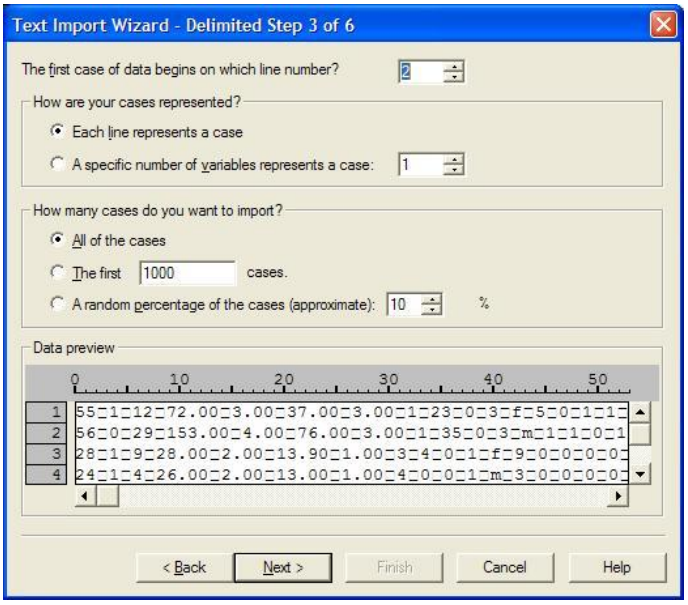

Figure 5.10: Text Import Wizard - Step 3 of 6.

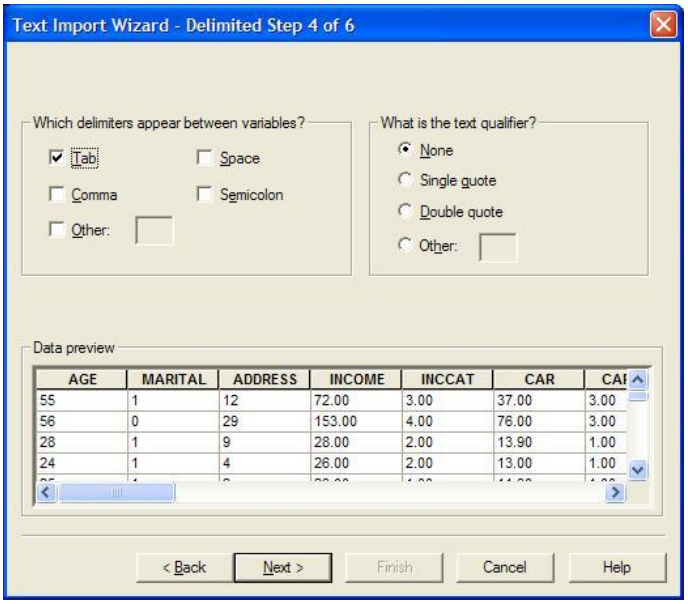

Figure 5.11: Text Import Wizard - Step 4 of 6.

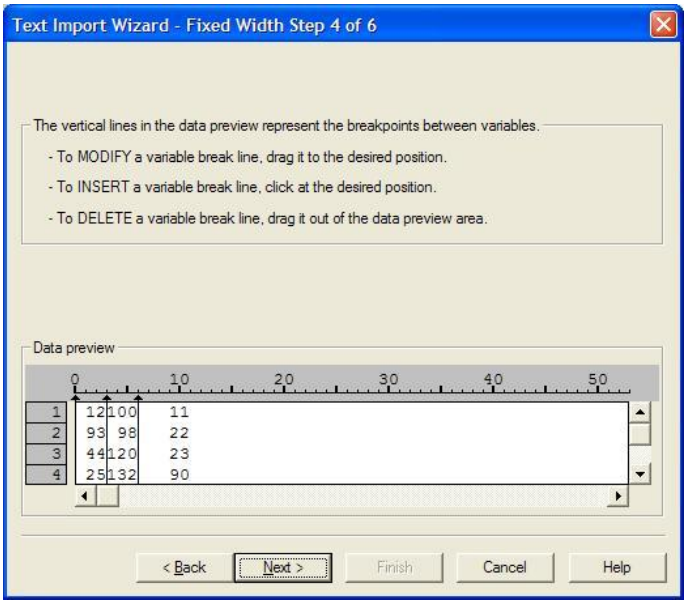

Figure 5.12: Text Import Wizard - Alternate Step 4 of 6.

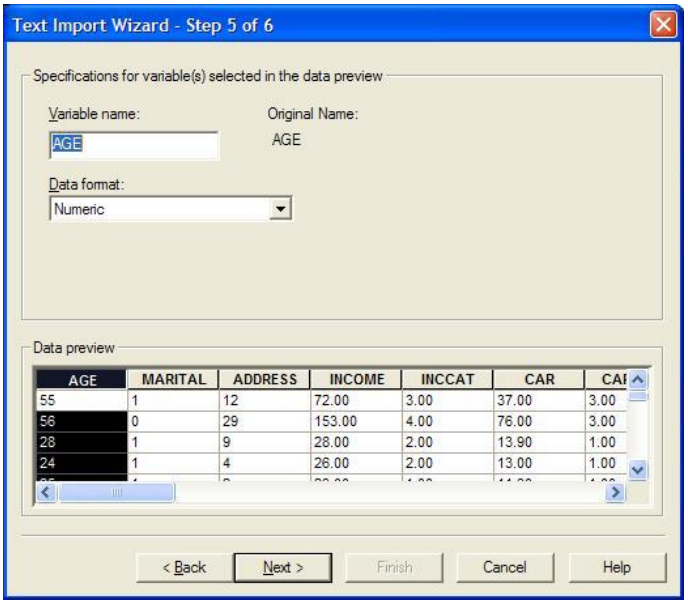

Figure 5.13: Text Import Wizard - Step 5 of 6.

In "Step 5 of 6" (5.13) you will have the chance to change the names of variables and/or the data format (numeric, data or string). Ordinarily you don't need to do anything at this step.

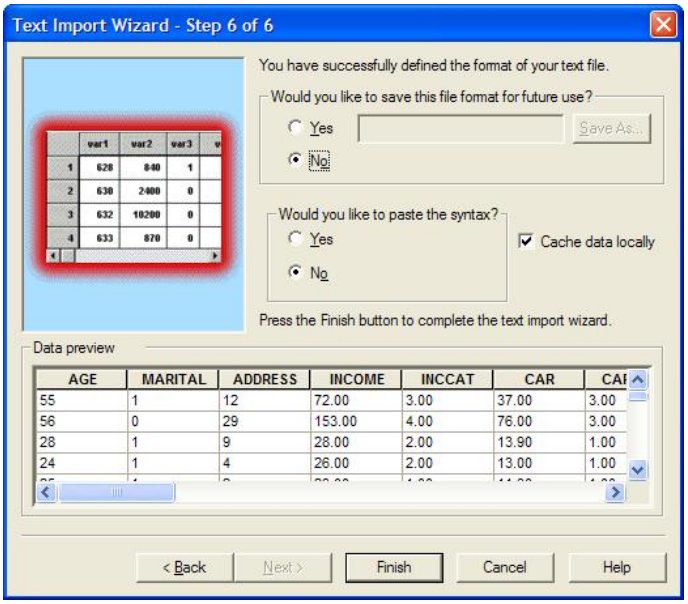

Figure 5.14: Text Import Wizard - Step 6 of 6.

In "Step 6 of 6" (5.14) you will have the chance to save all of your previous choices to simplify future loading of a similar file. We won't use this feature in this course, so you can just click the Finish button.

The most common error in loading data is forgetting to specify the presence of column headers in step 2. In that case the column header (variable names) appear as data rather than variable names.

5.5 Creating new variables

Creating new variables (data transformation) is commonly needed, and can be somewhat complicated. Depending on what you are trying to do, one of several menu options starts the process.

For creating of a simple data **transformation**, which is the result of applying a mathematical formula to one or more existing variables, use the ComputeVariable

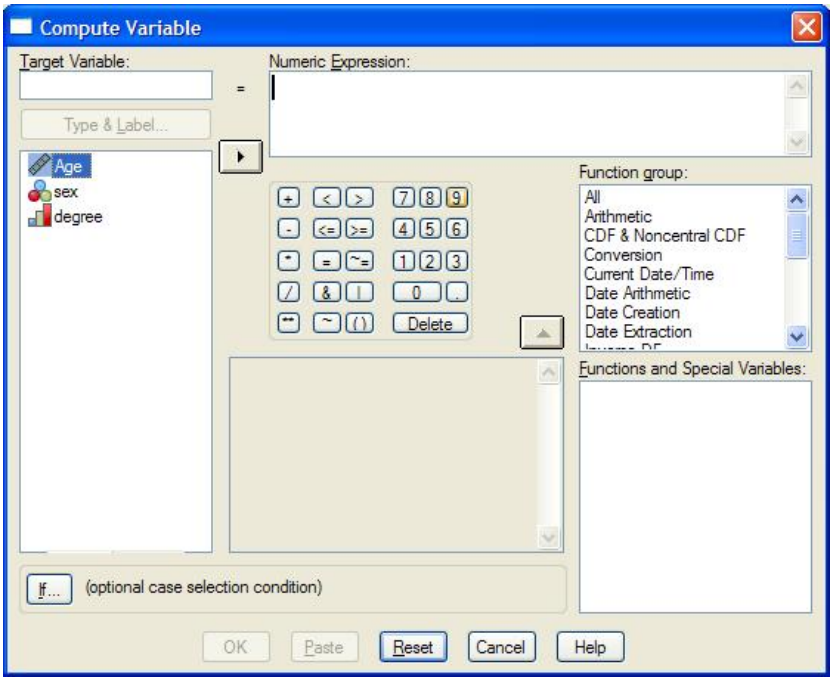

Figure 5.15: Compute Variable dialog box.

item on the Transform menu of the Data Editor. This open the Compute Variable dialog box (5.15). First enter a new variable name in the Target Variable box (remembering the naming rules discussed above). Usually you will want to click the "Type & Label" box to open another dialog box which allows you to enter a longer, more readable Label for the variable. (You will almost never want to change the type to "String".) Click Continue after entering the Label. Next you will enter the "Numeric Expression" in the Compute Variable dialog box. Two typical expressions are "log(weight)" which creates the new variable by the taking the log of the existing variable "weight", and "weight/height**2" which computes the body mass index from height and weight by dividing weight by the square (second power) of the height. (Don't enter the quotation marks.)

To create a transformation, use whatever method you can to get the required Numeric Expression into the box. You can either type a variable name or double click it in the variable list to the left, or single click it and click the right arrow. Spaces don't matter (except within variable names), and standard order of operations are used, but can be overridden with parentheses as needed. Numbers, operators (including * for times), and function names can be entered by clicking the mouse, but direct typing is usually faster. In addition to the help system, the list of functions may be helpful for finding the spelling of a function, e.g., sqrt for square root.

Comparison operators (such as =, <. and >) can be used with the understanding that the result of any comparison is either "true", coded as 1, or "false", coded as 0. E.g., if one variable called "vfee" has numbers indicating the size of a fee, and a variable called "surcharge" is 0 for no surcharge and 1 for a $25 surcharge, then we could create a new variable called "total" with the expression "vfee+25*(surcharge=1)". In that case either 25 (25*1) or 0 (25*0) is added to "vfee" depending of the value of "surcharge".

Advanced: To transform only some cases and leave others as "missing data" use the "If" button to specify an expression that is true only for the cases that need to be transformed.

Some other functions worth knowing about are ln, exp, missing, mean, min, max, rnd, and sum. The function ln() takes the natural log, as opposed to log(), which is common log. The function exp() is the anti-log of the natural log, as opposed to 10**x which is the common log's anti-log. The function missing() returns 1 if the variable has missing data for the case in question or 0 otherwise. The functions min(), max(), mean() and sum(), used with several variables separated with

commas inside the parentheses, computes a new value for each case from several existing variables for that case. The function rnd() rounds to a whole number.

5.5.1 Recoding

In addition to simple transformations, we often need to create a new variable that is a **recoding** of an old variable. This is usually used either to "collapse" categories in a categorical variable or to create a categorical version of a quantitative variable by "binning". Although it is possible to over-write the existing variable with the new one, I strongly suggest that you always preserve the old variable (for record keeping and in case you make an error in the encoding), and therefore you should use the 'into Different Variables" item under "Recode" on the "Transform" menu, which opens the "Recode into Different Variables" dialog box (5.16).

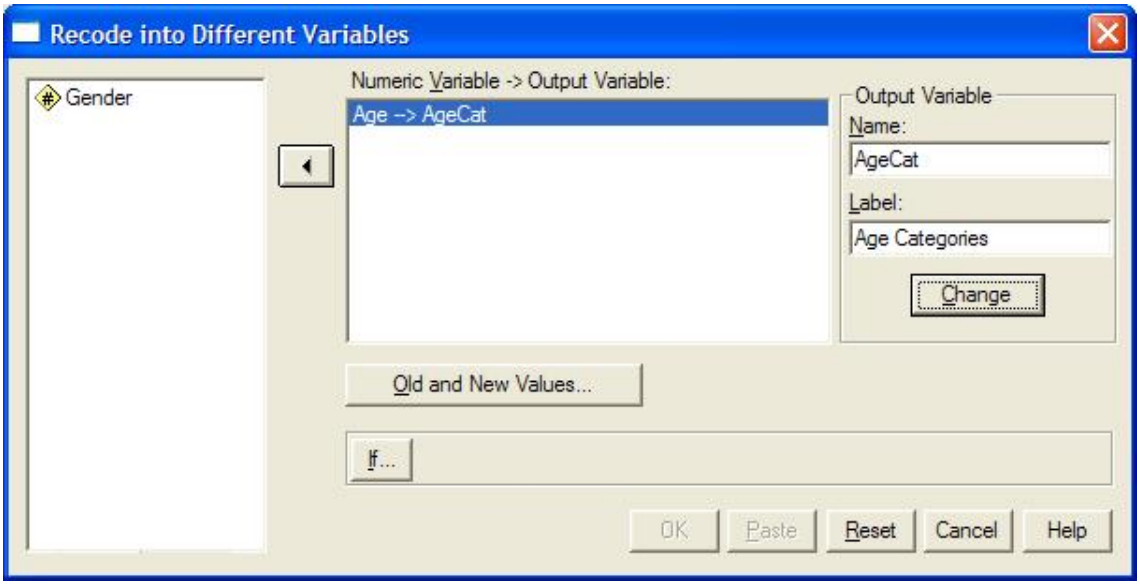

Figure 5.16: Recode into Different Variables Dialog Box.

First enter the existing variable name into the "Numeric Variable -> Output Variable" box. If you have several variables that need the same recoding scheme, enter each of them before proceeding. Then, for each existing variable, go to the "Output Variable" box and enter a variable Name and Label for the new recoded variable, and confirm the entry with the Change button.

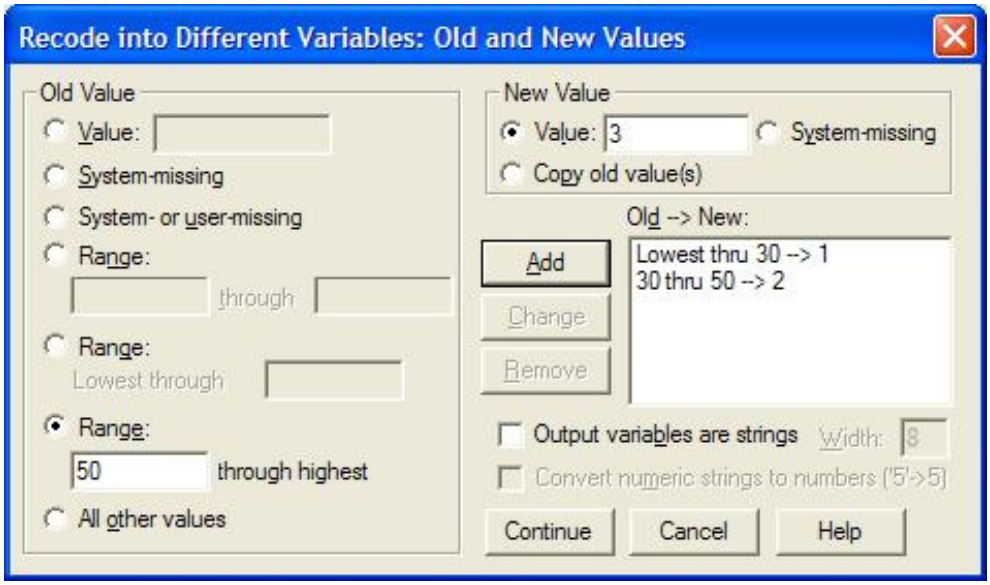

Figure 5.17: Recode into Different Variables: Old and New Values Dialog Box.

Then click the "Old and New Values" button to open the "Recode into Different Variables: Old and New Values" dialog box (5.17). Your goal is to specify as many "rules" as needed to create a new value for every possible old value so that the "Old–>New" box is complete and correct. For each one or several old values that will be recoded to a particular new value, enter the value or range of values on the left side of the dialog box, then enter the new value that represents the recoding of the old value(s) in the "New Value" box. Click Add to register each particular recoding, and repeat until finished. Often the "All other value" choice is the last choice for the "Old value". You can also use the Change and Remove buttons as needed to get a final correct "Old–>New" box. Click Continue to finalize the coding scheme and return to the "Recode into Different Values" box. Then click OK to create the new variable(s). If you want to go directly on to recode another variable, I strongly suggest that you click the Reset button first to avoid confusion.

5.5.2 Automatic recoding

Automatic recode is used in SPSS when you have strings (words) as the actual data levels and you want to convert to numbers (usually with Value labels). Among other reasons, this conversion saves computer memory space.

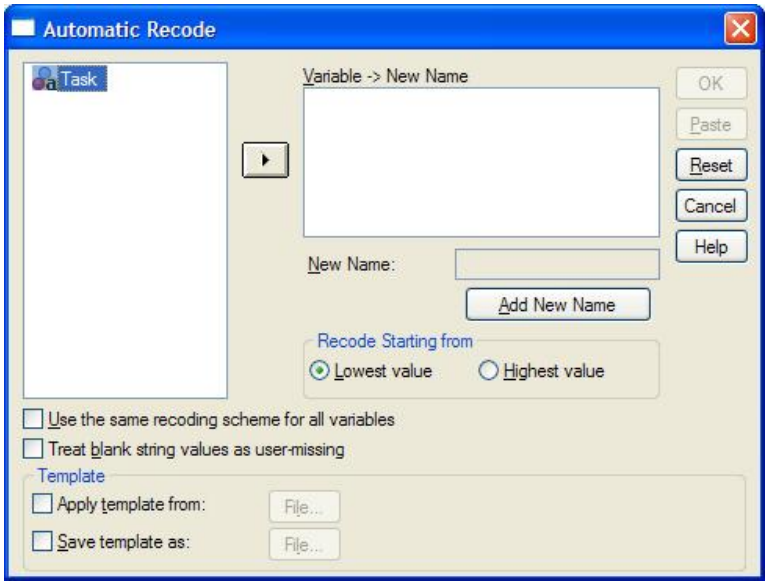

Figure 5.18: Automatic Recode Dialog Box.

From the Transform menu of the Data Editor menu, select "Automatic Recode" to get the "Automatic Recode" dialog box as shown in figure 5.18. Choose a variable, enter a new variable name in the "New Name" box and click "Add New Name". Repeat if desired for more variables. If there are missing data values in the variable and they are coded as blanks, click "Treat blank string values as user-missing". Click OK to create the new variable. You will get some output in the Output window showing the recoding scheme. A new variable will appear in the Data Window. If you click the Value Labels toolbar button, you will see that the new variable is really numeric with automatically created value labels.

5.5.3 Visual binning

SPSS has a option called "Visual Binning", accessed through the Visual Binning item on the Transformation menu, which allows you to interactively choose how to create a categorical variable from a quantitative (scale) variable. In the "Visual Binning" dialog box you select one or more quantitative (or ordinal) variables to work with, then click Continue. The next dialog box is also called "Visual Binning" and is shown in figure 5.19. Here you select a variable from the one(s) you previously chose, then enter a new name for the categorical variable you want

to create in the "Binned Variable" box (and optionally change its Label). A histogram of the variable appears. Now you have several choices for creating the "bins" that define the categories. One choice is to enter numbers in the Value column (and optionally Labels). For the example in the figure, I entered 33 as Value for line 1 and 50 for line 2, and the computer entered HIGH for line 3. I also entered the labels. When I click "OK" the quantitative variable "Age" will be recoded into a three level categorical variable based on my cutpoints.

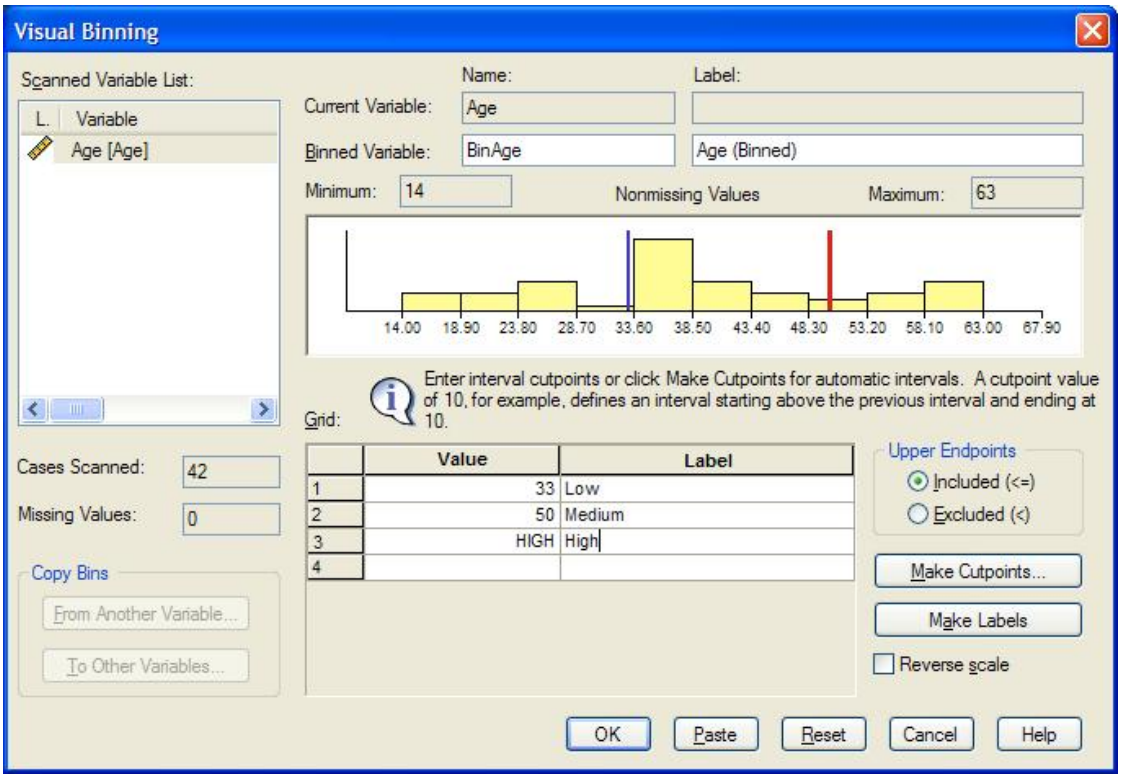

Figure 5.19: Visual Binning dialog box: Entered interval cutpoints.

The alternative to directly entering interval cutpoints is to click "Make Cutpoints" to open the "Make Cutpoints" dialog box shown in figure 5.20. Here your choices are to define some equal width intervals, equal percent intervals, or make cutpoints at fixed standard deviation intervals around the mean. After defining your cutpoints, click Apply to return to the histogram, which is now annotated based on your definition. (If you don't like the cutpoints edit them manually or return to Make Cutpoints.) You should manually enter meaningful labels for

the bins you have chosen or click "Make Labels" to get some computer generated labels. Then click OK to make your new variable.

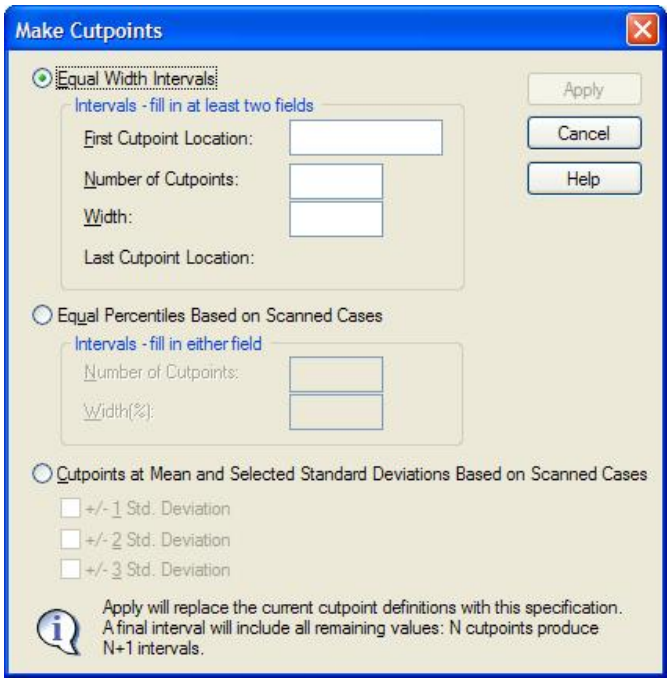

Figure 5.20: Visual Binning dialog box: Make cutpoints.

5.6 Non-graphical EDA

To **tabulate a single categorical variable**, i.e., get the numbers and percent of cases at each level of the variable, use the Frequencies subitem under the Descriptive Statistics item of the Analyze menu. This is also useful for quantitative variables with not too many unique values. When you choose your variable(s) and click OK, the Frequency table will appear in the Output Window. The default output (e.g., figure 5.21) shows each unique value, and its frequency and percent. The "Valid Percent" column calculates percents for only the non-missing data, while the "Percent" column only adds to 100% when you include the percent missing. Cumulative Percent can be useful for ordinal data. It adds all of the Valid Percent numbers for any row plus all rows above in the table, i.e. for any data value it shows what percent of cases are less than or equal to that value.

degree

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	14	33.3	33.3	33.3
	2	15	35.7	35.7	69.0
	3	13	31.0	31.0	100.0
	Total	42	100.0	100.0	

Figure 5.21: SPSS frequency table.

To **cross-tabulate two or more categorical variables** use the Crosstabs subitem under the Descriptive Statistics item of the Analyze menu. This is also useful for quantitative variables with not too many unique values. Enter one variable under "Rows" and one under "Columns". If you have a third variable, enter it under "Layer". (You can use the "Next" Layer button if you have more than three variables to cross-tabulate, but that may be too hard to interpret. Click OK to get the cross-tabulation of the variables. The default is to show only the counts for each combination of levels of the variables. If you want percents, click the "Cells" button before clicking OK; this gives the "Crosstabs: Cell Display" dialog box from which you can select percentages that add to 100% across each Row, down each "Column" or in "Total" across the whole cross-tabulation. Try to think about which of these makes the most sense for understanding your dataset it each particular case. Example output is shown in figure 5.22.

sex * degree Crosstabulation

			degree			Total
			1	2	3	
sex	Male	Count	7	5	8	20
		% within degree	50.0%	33.3%	61.5%	47.6%
	Female	Count	7	10	5	22
		% within degree	50.0%	66.7%	38.5%	52.4%
Total		Count	14	15	13	42
		% within degree	100.0%	100.0%	100.0%	100.0%

Figure 5.22: SPSS cross-tabulation.

For various **univariate quantitative variable sample statistics** use the

Descriptives subitem under the Descriptive Statistics item of the Analyze menu. Ordinarily you should use "Descriptives" for quantitative and possibly ordinal variables. (It words, but rarely makes sense for nominal variables.) The default is to calculate the sample mean, sample "Std. deviation", sample minimum and sample maximum. You can click on "Options" to access other sample statistics such as sum, variance, range, kurtosis, skewness, and standard error of the mean. Example output is show in figure 5.23. The sample size (and indication of any missing values) is always given. Note that for skewness and kurtosis standard errors are given. The rough rule-of-thumb for interpreting the skewness and kurtosis statistics is to see if the absolute value of the statistic is smaller than twice the standard error (labeled Std. Error) of the corresponding statistic. If so, there is no good evidence of skewness (asymmetry) or kurtosis. If the absolute value is large (compared to twice the standard error), then a positive number indicates right skew or positive kurtosis respectively, and a negative number indicates left skew or negative kurtosis.

Rule of thumb: Interpret skewness and kurtosis sample statistics by comparing the absolute value of the statistic to twice the standard error of the statistic. Small statistic value are consistent with the zero skew and kurtosis of a Gaussian distribution.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	41	14	63	38.59	13.240
degree	42	1	3	1.98	.811
Valid N (listwise)	41				

Figure 5.23: SPSS descriptive statistics.

To get the **correlation of two quantitative variables** in SPSS, from the Analyze menu item choose Correlate/Bivariate. Enter two (or more) quantitative variables into the Variables box, then click OK. The output will show correlations and a p-value for the test of zero correlation for each pair of variables. You may also want to turn on calculation of means and standard deviations using the Options button.Example output is show in figure 5.24. The "Pearson Correlation" statis-

tic is the one that best estimates the population correlation of two quantitative variables discussed in section 3.5.

Correlations

		Age	Response time (ms)
Age	Pearson Correlation	1.000	-.252
	Sig. (2-tailed)		.283
	N	20.000	20
Response time (ms)	Pearson Correlation	-.252	1.000
	Sig. (2-tailed)	.283	
	N	20	20.000

Figure 5.24: SPSS correlation.

(To calculate the various types of correlation for categorical variables, run the crosstabs, but click on the "Statistics" button and check "Correlations".)

To calculate **median or quartiles for a quantitative variable** (or possibly an ordinal variable) use Analyze/Frequencies (which is normally used just for categorical data), click the Statistics button, and click median and/or quartiles. Normally you would also uncheck "Display frequency tables" in the main Frequencies dialog box to avoid voluminous, unenlightening output. Example output is show in figure 5.25.

Statistics

degree

N	Valid	42
	Missing	0
Median		2.00
Percentiles	25	1.00
	50	2.00
	75	3.00

Figure 5.25: SPSS median and quantiles.

5.7 Graphical EDA

5.7.1 Overview of SPSS Graphs

The Graphs menu item in SPSS version 16.0 has two sub-items: ChartBuilder and LegacyDialogs. As you might guess, the legacy dialogs item access older ways to create graphs. Here we will focus on the interactive Chart Builder approach. Note that graph, chart, and plot are interchangeable terms.

There is a great deal of flexibility in building graphs, so only the principles are given here.

When you select the Chart Builder menu item, it will bring up the Chart Builder dialog box. Note the three main areas: the variable box at top left, the chart preview area (also called the "canvas") at top right, and the (unnamed) lower area from which you can select a tab out of this group of tabs: Gallery, Basic Elements, Groups/PointID, and Titles/Footnotes.

A view of the (empty) Chart Builder is shown in 5.26.

To create a graph, go to the Gallery tab, select a graph type on the left, then choose a suitable template on the right, i.e. one that looks roughly like the graph you want to create. Note that the templates have names that appear as pop-up labels if you hover the mouse over them. Drag the appropriate template onto the canvas at top right. A preview of your graph (but not based on your actual data) will appear on the canvas.

The use of the Basic Elements tab is beyond the scope of this chapter.

The Groups/PointsID tab (5.27) serves both to add additional information from auxiliary variables (Groups) and to aid in labeling outliers or other interesting points (Point ID). After placing your template on the canvas, select the Groups/PointID tab. Sex check boxes are present in this tab. The top five choices refer to grouping, but only the ones appropriate for the chosen plot will be active. Check whichever ones might be appropriate. For each checked box, a "drop zone" will be added to the canvas, and adding an auxiliary variable into the drop zone (see below) will, in some way that is particular to the kind of graph you are creating, cause the graphing to be split into groups based on each level of the auxiliary variable. The "Point ID label" check box (where appropriate) adds a drop zone which hold the name of the variable that you want to use to label outliers or other special points. (If you don't set this, the row number in the spreadsheet is used

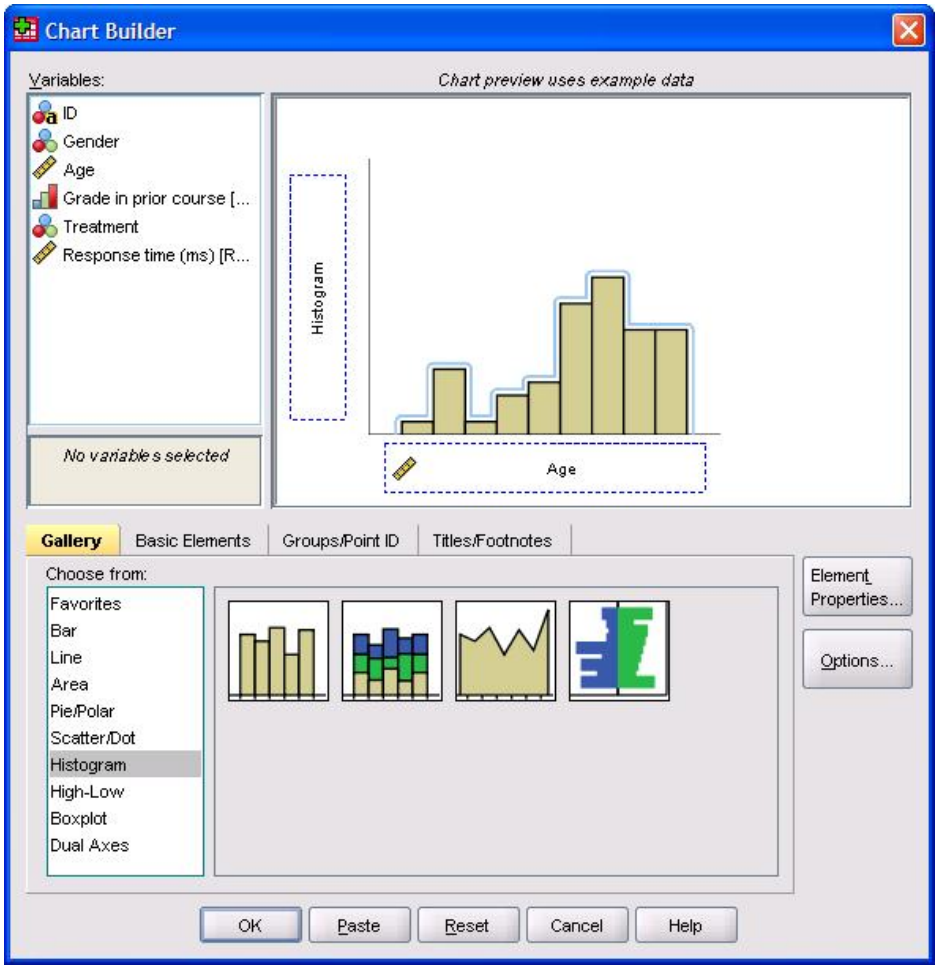

Figure 5.26: SPSS Empty Chart Builder.

for labeling.)

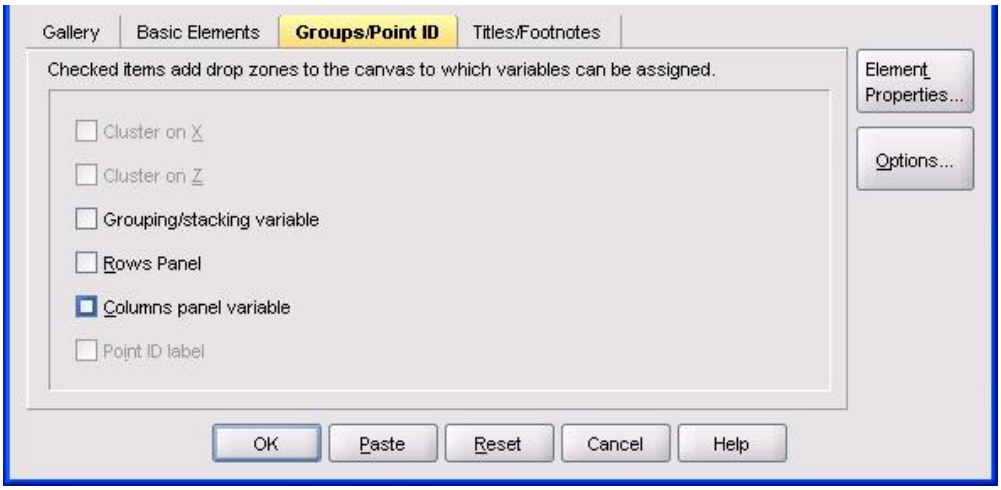

Figure 5.27: SPSS Groups/Point ID tab of Chart Builder.

The Titles/Footnotes tab (5.28) has check boxes for titles and footnotes. Check any that you need to appropriately annotate your graph. When you do so, the Element Properties dialog box (5.29) will open. (You can also open and close this box with the Element Properties button.) In the Element Properties box, select each title and/or footnote, then enter the desired annotation in the "Content" box.

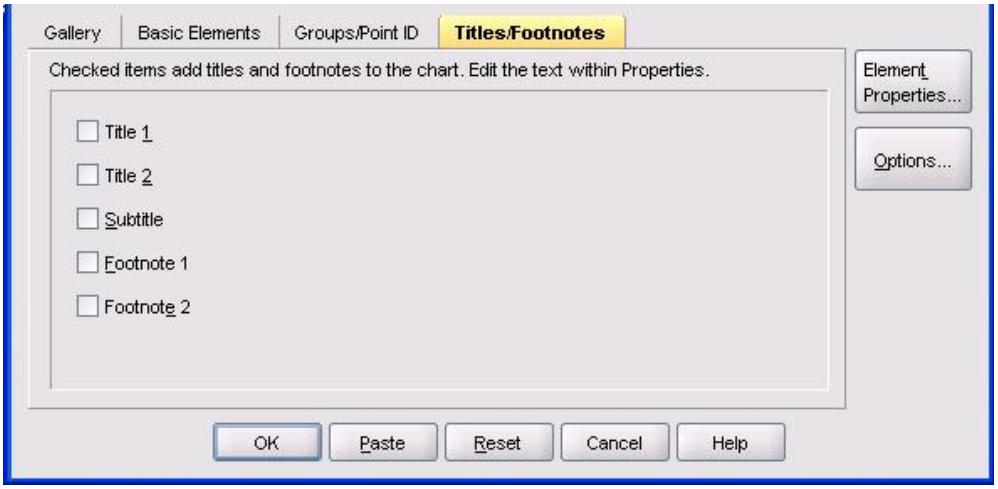

Figure 5.28: SPSS Titles/Footnote tab of Chart Builder.

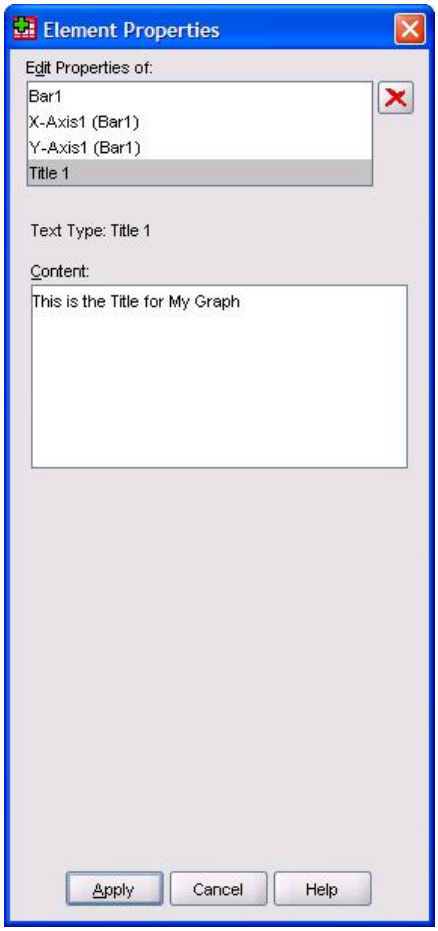

Figure 5.29: SPSS Element Properties dialog box.

Next you will add all of the variables that participate in the production of your graph to the appropriate places on the canvas. Note that when you click on any categorical variable in the Variables box, its categories are listed below the variable box. Drag appropriate variables into the pre-specified drop boxes (which vary with the type of graph chosen, and may include things like the x-axis and y-axis), as well as the drop boxes you created from the Groups/PointID tab.

You may want to revisit the Element Properties box and click through each element of the "Edit Properties of" box to see if there are any properties you might want to alter (e.g., the order of appearance of the levels of a categorical variable, or the scale for a quantitative variable). Be sure to click the Apply button after making any changes and before selecting another element or closing the Element Properties box.

Finally click OK in the Chart Builder dialog box to create your plot. It will appear at the end of your results in the SPSS Viewer window.

When you re-enter the Chart Builder, the old information will still be there, and that is useful to tweak the appearance of a plot. If you want to create a new plot unrelated to the previous plot, you will probably find it easiest to use the Reset button to remove all of the old information.

5.7.2 Histogram

The basic univariate **histogram** for quantitative or categorical data is generated by using the Simple Histogram template, which is the first one under Histogram in the Gallery. Simply drag your variable onto the x-axis to define your histogram ("Histogram" will appear on the y-axis.). For optionally grouping by a second variable, check "Grouping/stacking variable" in the Groups/PointID tab, then drag the second variable to the "Stack:set color" drop box. The latter is equivalent to choosing the "Stacked Histogram" in the gallery.

A view of the Chart Builder after setting up a histogram is shown in 5.30.

The "Population Pyramid" template (on the right side of the set of Histogram templates) is a nice way to display histograms of one variable at all levels of another (categorical) variable.

To **change the binning of a histogram**, double click on the histogram in the SPSS Viewer, which opens the Chart Editor (5.31), then double click on a histogram bar in the Chart Editor to open the Properties dialog box (5.32). Be

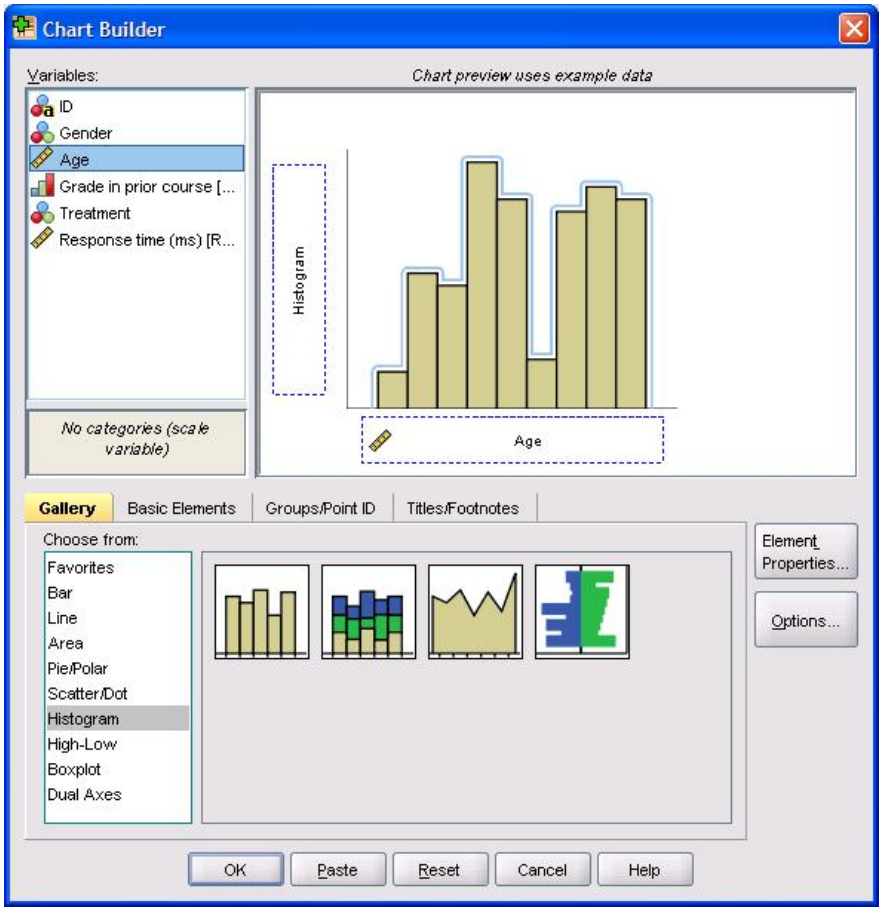

Figure 5.30: SPSS histogram setup.

sure that the Binning tab is active. Under "X Axis" change from Automatic to Custom, then enter either the desired number of intervals of the desired interval width. Click apply to see the result. When you achieve the best result, click Close in the Properties window, then close the Chart Editor window.

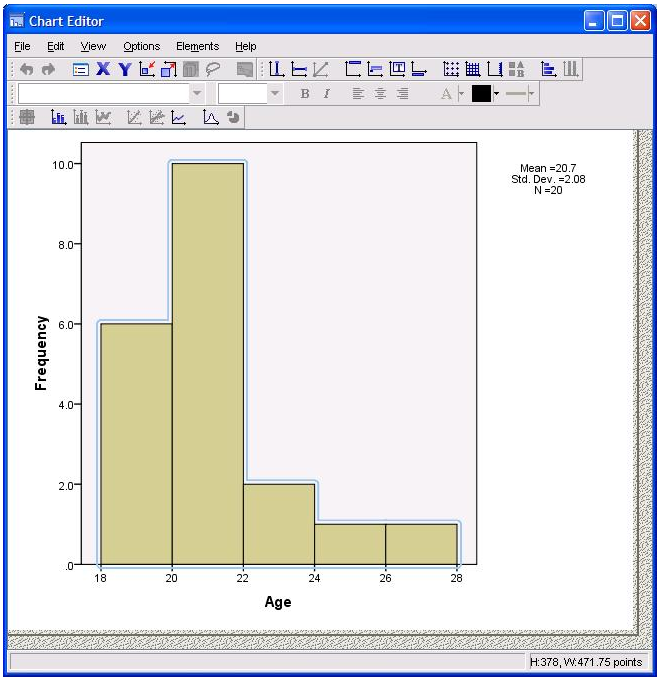

Figure 5.31: SPSS Chart Editor.

An example of a histogram produced in SPSS is shown in figure 5.33.

For histograms or any other graphs, it is a good idea to use the Titles/Footnote tab to set an appropriate title, subtitle and/or footnote.

5.7.3 Boxplot

A **boxplot** for quantitative random variables is generated in SPSS by using one of the three boxplot templates in the Gallery (called simple, clustered, and 1-D, from left to right). The 1-D boxplot shows the distribution of a single variable. The simple boxplot shows the distribution a one (quantitative) variable at each level of another (categorical) variable. The clustered boxplot shows the distribution a one (quantitative) variable at each level of two other categorical variables.

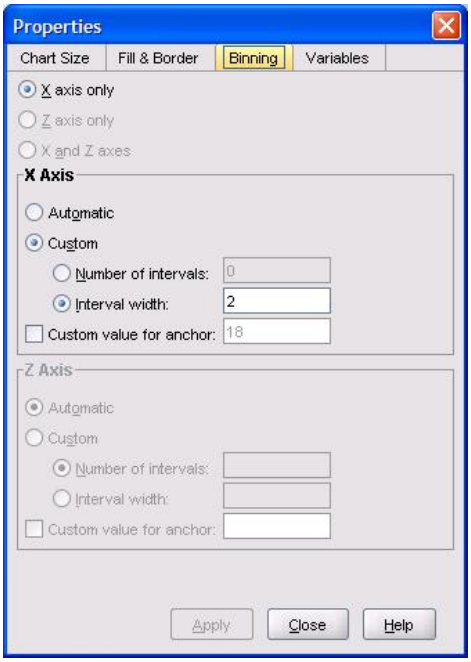

Figure 5.32: Binning in the SPSS Chart Editor.

An example of the Chart Builder setup for a simple boxplot with ID labels is shown in figure 5.34. The corresponding plot is in figure 5.35.

Other univariate graphs, such as pie charts and bar charts are also available through the Chart Builder Gallery.

5.7.4 Scatterplot

A **scatterplot** is the best EDA for examining the relationship between two quantitative variables, with a "point" on the plot for each subject. It is constructed using templates from the Scatter/Dot section of the Chart Builder Gallery. The most useful ones are the first two: Simple Scatter and Grouped Scatter. Grouped Scatter adds the ability to show additional information from some categorical variable, in the form of color or symbol shape.

Once you have placed the template on the canvas, drag the appropriate quantitative variables onto the x- and y-axes. *If one variable is outcome and the other explanatory, be sure to put the outcome on the vertical axis.* A simple example is

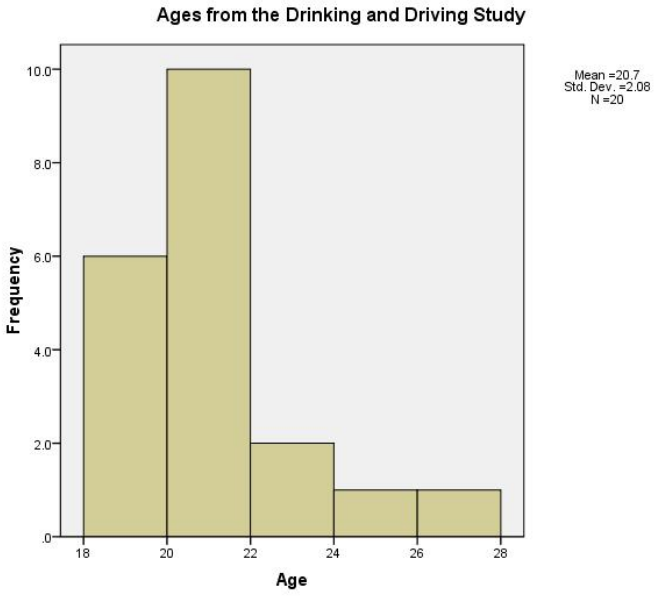

Figure 5.33: SPSS histogram.

shown in figure 5.36. The corresponding plot is in figure 5.37.

You can further modify a scatter plot by adding a best-fit straight line or a "non-parametric" smooth curve. This is done using the Chart Editor rather than the Chart Builder, so it is an addition to a scatterplot already created. Open the Chart Editor by double clicking on the scatterplot in the SPSS Viewer window. Choose "Add Fit Line at Total" by clicking on the toolbar button that looks like a scatterplot with a fit line through it, or by using the menu option Elements/FitLineAtTotal. This brings up the a Properties box with a "Fit Line" tab (5.38). The "Linear" Fit Method adds the best fit linear regression line. The "Loess" Fit Method adds a "smoother" line to your scatterplot. The smoother line is useful for detecting whether there is a non-linear relationship. (Technically it is a kernel smoother.) There is a degree of subjectivity in the overall smoothness vs. wiggliness of the smoother line, and you can adjust the "% of points to fit" to change this. Also note that if you have groups defined with separate point colors for each group, you can substitute "Add Fit Line at Subgroups" for "Add Fit Line at Total" to have separate lines for each subgroup.

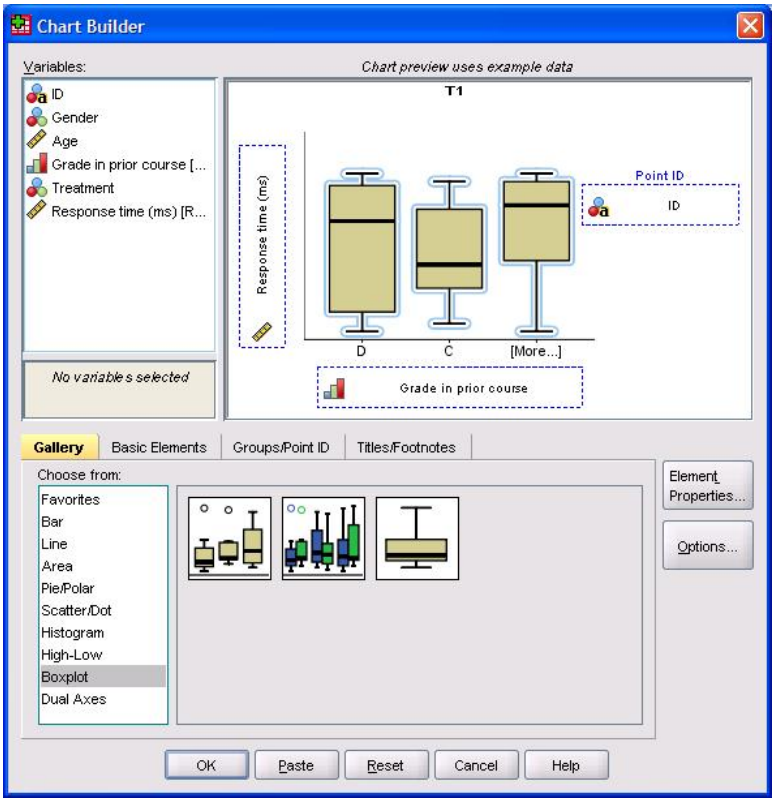

Figure 5.34: SPSS boxplot setup in Chart Builder.

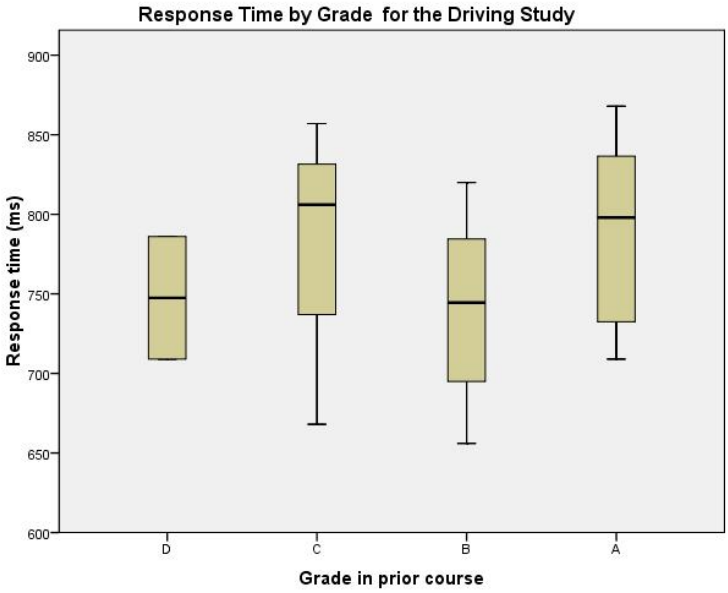

Figure 5.35: SPSS boxplot.

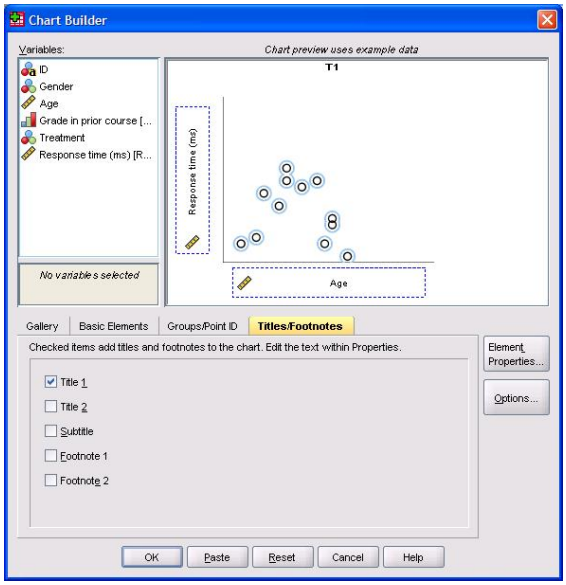

Figure 5.36: SPSS scatterplot setup in Chart Builder.

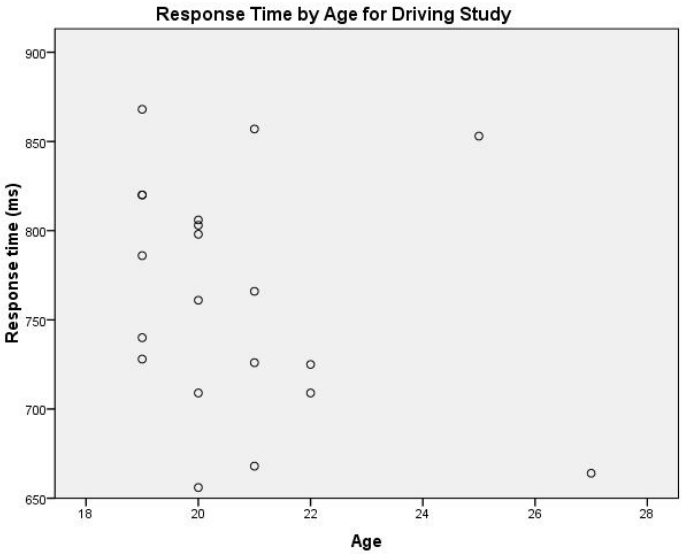

Figure 5.37: SPSS simple scatterplot.

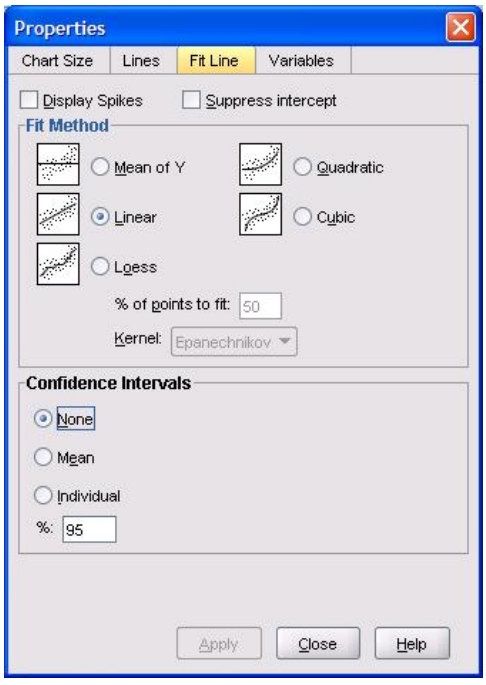

Figure 5.38: SPSS Fit Line tab of Chart Editor.

5.8 SPSS convenience item: Explore

The Analyze/DescriptiveStatistics/Explore menu item in SPSS is a convenience menu item that performs several reasonable EDA steps, both graphical and non-graphical for a quantitative outcome and a categorical explanatory variable (factor). "Explore" is *not* a standard statistical term; it is only an SPSS menu item. So don't use the term in any formal setting!

In the Explore dialog box you can enter one or more quantitative variables in the "Dependent List" box and one or more categorical variables in the "Factor List" box. For each variable in the "Factor List", a complete section of output will be produced. Each section of output examines each of the variables on the "Dependent List" separately. For each outcome variable, graphical and non-graphical EDA are produced that examine the outcome broken down into groups determined by the levels of the "factor". A partial example is given in figure 5.39. In addition to the output shown in the figure, stem-and-leaf plots and side-by-side boxplots are produced by default. The choice of plots and statistics can be changed in the Explore dialog box.

This example has "strength" as the outcome and "sex" as the explanatory variable (factor). The "Case Processing Summary" tells us the number of cases and information about missing data separately for each level of the explanatory variable. The "Descriptives" section gives a variety of statistics for the strength outcome broken down separately for males and females. These statistics include mean and confidence interval on the mean (i.e., the range of means for which we are 95% confident that the true population mean parameter falls in). (The CI is constructed using the "Std. Error" of the mean.) Most of the other statistics should be familiar to you except for the "5% trimmed mean"; this is a "robust" measure of central tendency equal to the mean of the data after throwing away the highest and lowest 5% of the data. As mentioned on page 125, standard errors are calculated for the sample skewness and kurtosis, and these can be used to judge whether the observed values are close or far from zero (which are the expected skewness and kurtosis values for Gaussian data).

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	sex	N	Percent	N	Percent	N	Percent
Strength	Male	20	100.0%	0	.0%	20	100.0%
	Female	21	95.5%	1	4.5%	22	100.0%

Descriptives

	sex			Statistic	Std. Error
Strength	Male	Mean		22.015	.5390
		95% Confidence Interval for Mean	Lower Bound	20.887	
			Upper Bound	23.143	
		5% Trimmed Mean		22.106	
		Median		21.700	
		Variance		5.810	
		Std. Deviation		2.4103	
		Minimum		16.5	
		Maximum		25.9	
		Range		9.4	
		Interquartile Range		3.1	
		Skewness		-.412	.512
		Kurtosis		.016	.992
	Female	Mean		22.119	.4502
		95% Confidence Interval for Mean	Lower Bound	21.180	
			Upper Bound	23.058	
		5% Trimmed Mean		22.191	
		Median		22.300	
		Variance		4.257	
		Std. Deviation		2.0632	
		Minimum		17.8	
		Maximum		25.1	
		Range		7.3	
		Interquartile Range		3.5	
		Skewness		-.408	.501
		Kurtosis		-.727	.972

Figure 5.39: SPSS "Explore" output.