IBM SPSS Statistics 25 Brief Guide

Note
Before using this information and the product it supports, read the information in "Notices" on page 93.

Product Information

This edition applies to version 25, release 0, modification 0 of IBM SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Introduction

This guide will show you how to use many of the available features. It is designed to provide a step-by-step, hands-on guide. All of the files shown in the examples are installed with the application so that you can follow along, performing the same analyses and obtaining the same results shown here.

If you want detailed examples of various statistical analysis techniques, try the step-by-step Case Studies, available from the Help menu.

Sample Files

Most of the examples that are presented here use the data file *demo.sav*. This data file is a fictitious survey of several thousand people, containing basic demographic and consumer information.

If you are using the Student version, your version of *demo.sav* is a representative sample of the original data file, reduced to meet the 1,500-case limit. Results that you obtain using that data file will differ from the results shown here.

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

Opening a Data File

To open a data file:

1. From the menus choose:

   File > **Open** > **Data...**

   A dialog box for opening files is displayed.

   By default, IBM® SPSS® Statistics data files (*.sav* extension) are displayed.

   This example uses the file *demo.sav*.

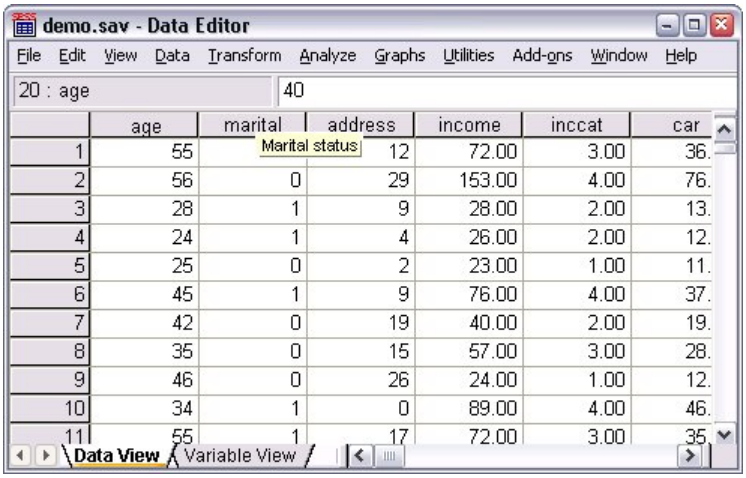


Figure 1. demo.sav file in Data Editor

The data file is displayed in the Data Editor. In Data View, if you put the mouse cursor on a variable name (the column headings), a more descriptive variable label is displayed (if a label has been defined for that variable).

By default, the actual data values are displayed. To display labels:

2. From the menus choose:

View > **Value Labels**

Figure 2. Value Labels button

Alternatively, you can use the Value Labels button on the toolbar.

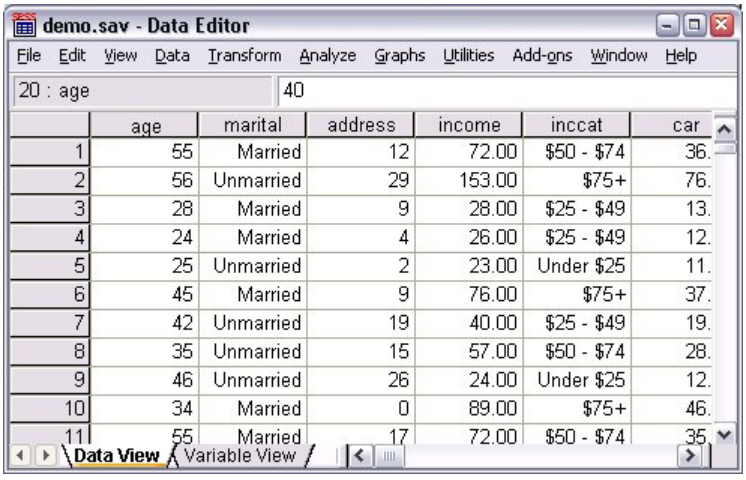


Figure 3. Value labels displayed in the Data Editor

Descriptive value labels are now displayed to make it easier to interpret the responses.

Running an Analysis

If you have any add-on options, the Analyze menu contains a list of reporting and statistical analysis categories.

We will start by creating a simple frequency table (table of counts). This example requires Statistics Base Edition.

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Frequencies...**

   The Frequencies dialog box is displayed.

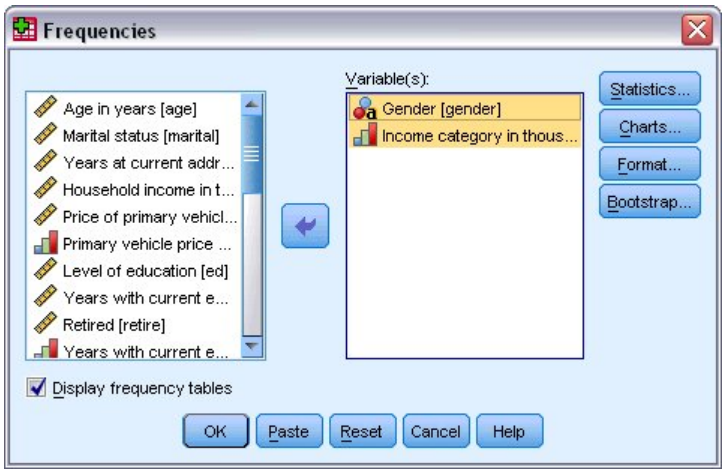


Figure 4. Frequencies dialog box

An icon next to each variable provides information about data type and level of measurement.

	Numeric	**String**	**Date**	**Time**
Scale (Continuous)		n/a		
Ordinal				
Nominal				

If the variable label and/or name appears truncated in the list, the complete label/name is displayed when the cursor is positioned over it. The variable name *inccat* is displayed in square brackets after the descriptive variable label. *Income category in thousands* is the variable label. If there were no variable label, only the variable name would appear in the list box.

You can resize dialog boxes just like windows, by clicking and dragging the outside borders or corners. For example, if you make the dialog box wider, the variable lists will also be wider.

In the dialog box, you choose the variables that you want to analyze from the source list on the left and drag and drop them into the Variable(s) list on the right. The **OK** button, which runs the analysis, is disabled until at least one variable is placed in the Variable(s) list.

In many dialogs, you can obtain additional information by right-clicking any variable name in the list and selecting **Variable Information** from the pop-up menu.

2. Click *Gender [gender]* in the source variable list and drag the variable into the target Variable(s) list.
3. Click *Income category in thousands [inccat]* in the source list and drag it to the target list.

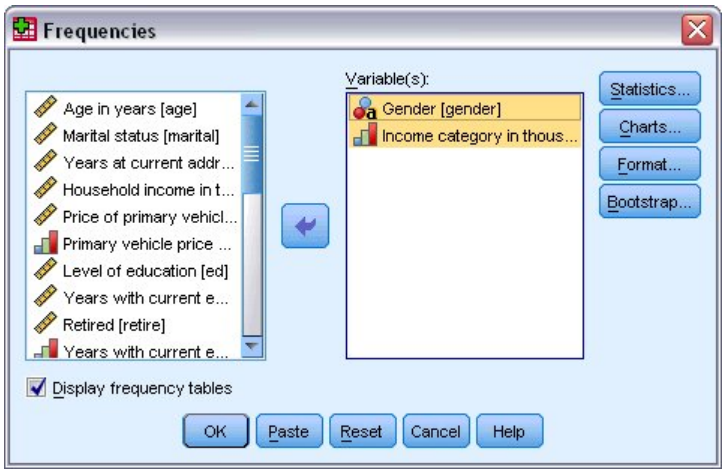


Figure 5. Variables selected for analysis

4. Click **OK** to run the procedure.

Results are displayed in the Viewer window.

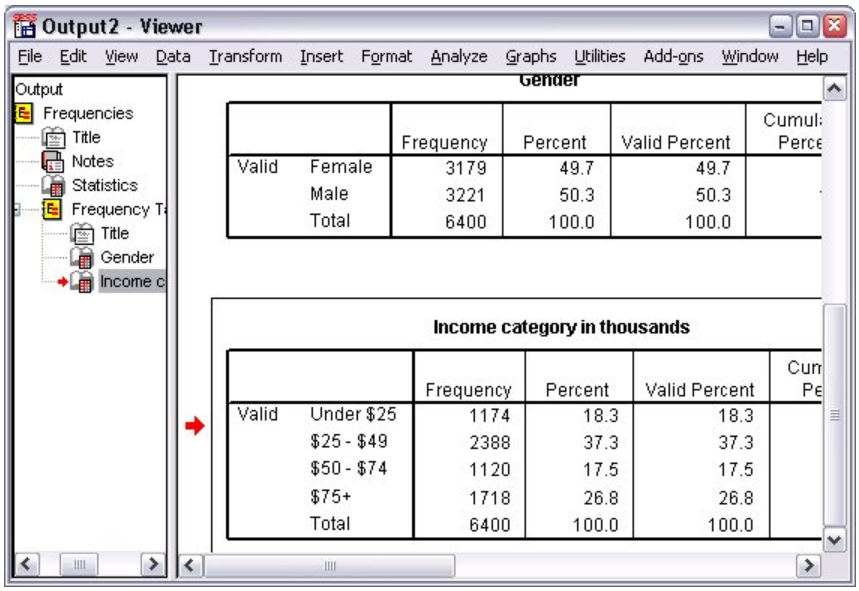


Figure 6. Frequency table of income categories

Creating Charts

Although some statistical procedures can create charts, you can also use the Graphs menu to create charts.

For example, you can create a chart that shows the relationship between wireless telephone service and PDA (personal digital assistant) ownership.

1. From the menus choose:
   Graphs > **Chart Builder...**

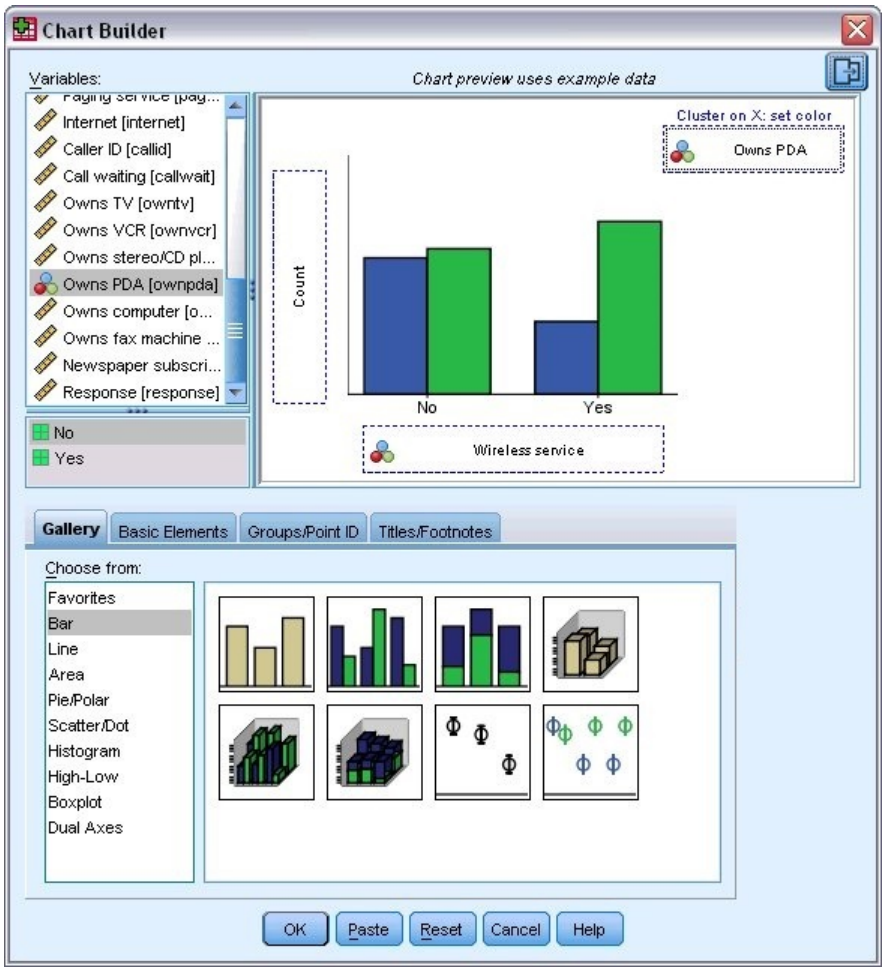


Figure 7. Chart Builder dialog box with completed drop zones

2. Click the **Gallery** tab (if it is not selected).
3. Click **Bar** (if it is not selected).
4. Drag the Clustered Bar icon onto the canvas, which is the large area above the Gallery.
5. Scroll down the Variables list, right-click *Wireless service [wireless]*, and then choose **Nominal** as its measurement level.
6. Drag the *Wireless service [wireless]* variable to the *x* axis.
7. Right-click *Owns PDA [ownpda]* and choose **Nominal** as its measurement level.
8. Drag the *Owns PDA [ownpda]* variable to the cluster drop zone in the upper right corner of the canvas.
9. Click **OK** to create the chart.

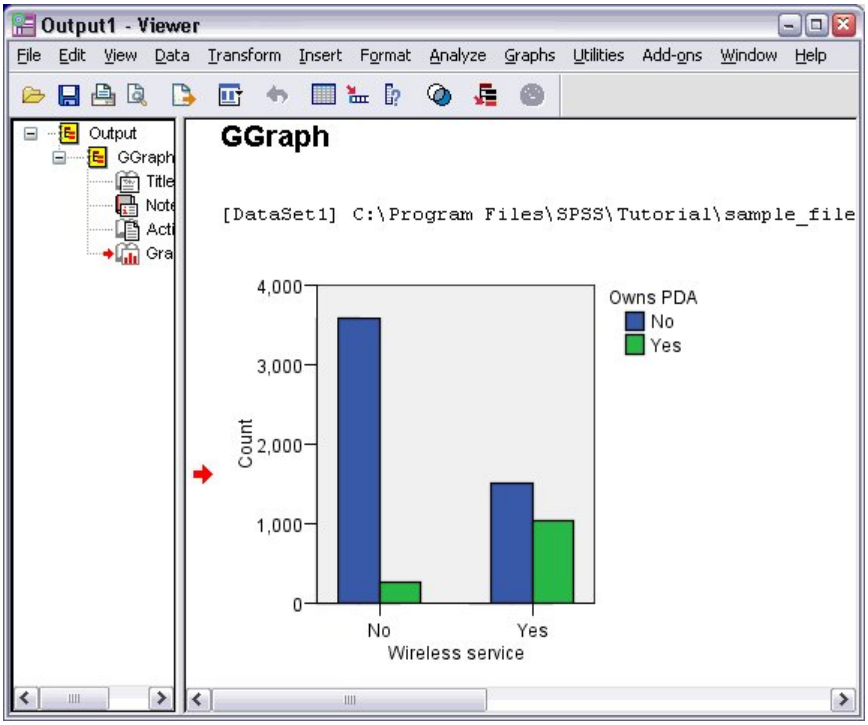


Figure 8. Bar chart displayed in Viewer window

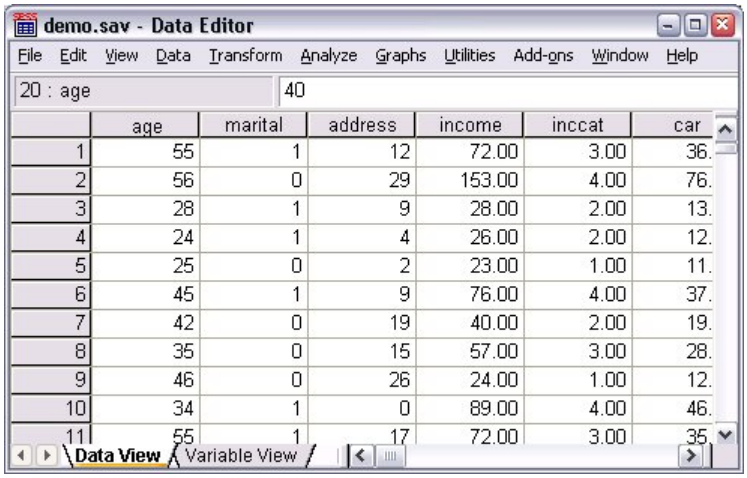
The bar chart is displayed in the Viewer. The chart shows that people with wireless phone service are far more likely to have PDAs than people without wireless service.

You can edit charts and tables by double-clicking them in the contents pane of the Viewer window, and you can copy and paste your results into other applications. Those topics will be covered later.

Chapter 2. Reading Data

Data can be entered directly, or it can be imported from a number of different sources. The processes for reading data stored in IBM SPSS Statistics data files; spreadsheet applications, such as Microsoft Excel; database applications, such as Microsoft Access; and text files are all discussed in this chapter.

Basic Structure of IBM SPSS Statistics Data Files



	age	marital	address	income	inccat	car
1	55	1	12	72.00	3.00	36.
2	56	0	29	153.00	4.00	76.
3	28	1	9	28.00	2.00	13.
4	24	1	4	26.00	2.00	12.
5	25	0	2	23.00	1.00	11.
6	45	1	9	76.00	4.00	37.
7	42	0	19	40.00	2.00	19.
8	35	0	15	57.00	3.00	28.
9	46	0	26	24.00	1.00	12.
10	34	1	0	89.00	4.00	46.
11	55	1	17	72.00	3.00	35.

Figure 9. Data Editor

IBM SPSS Statistics data files are organized by cases (rows) and variables (columns). In this data file, cases represent individual respondents to a survey. Variables represent responses to each question asked in the survey.

Reading IBM SPSS Statistics Data Files

IBM SPSS Statistics data files, which have a *.sav* file extension, contain your saved data.

1. From the menus choose:

   File > **Open** > **Data...**
2. Browse to and open *demo.sav*. See the topic Chapter 10, “Sample Files,” on page 83 for more information.

The data are now displayed in the Data Editor.

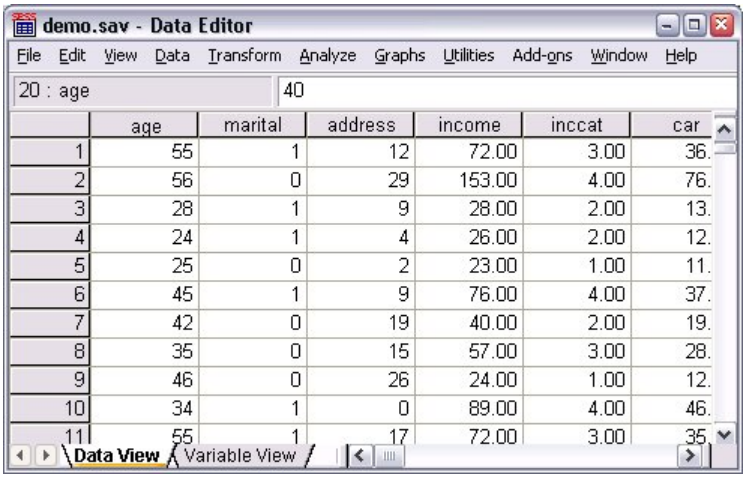


Figure 10. Opened data file

Reading Excel Data

Rather than typing all of your data directly into the Data Editor, you can read data from applications such as Microsoft Excel. You can also read column headings as variable names.

1. From the menus choose:

   File > **Import Data** > **Excel**

2. Go to the `Samples\English` folder and select `demo.xlsx`.

   The Read Excel File dialog displays a preview of the data file. The contents of the first sheet in the file are displayed. If the file has multiple sheets, you can select the sheet from the list.

   You can see that some of the string values for *Gender* have leading spaces. Some of the values for *MaritalStatus* are displayed as periods (.).

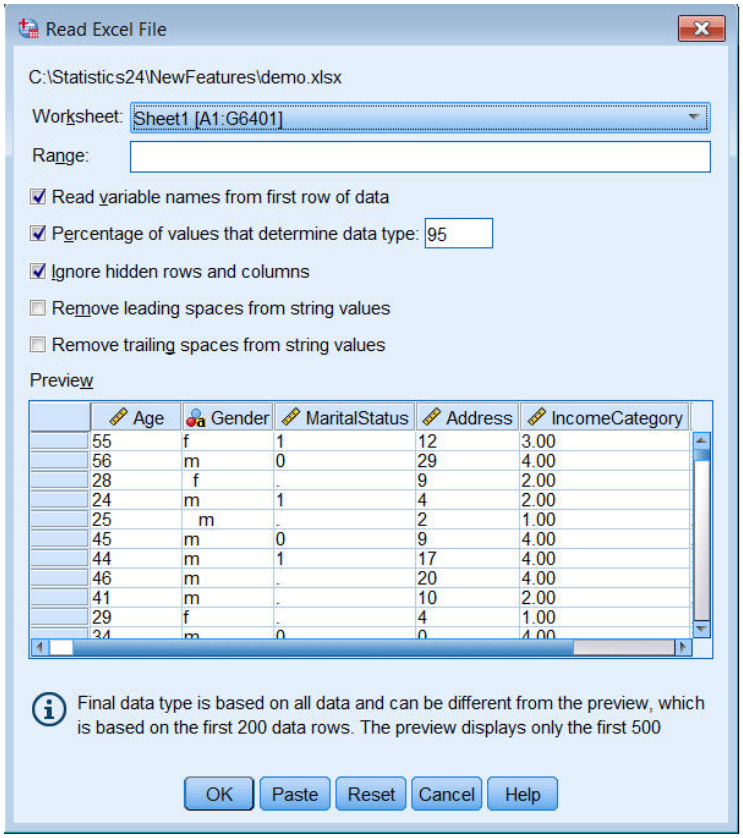
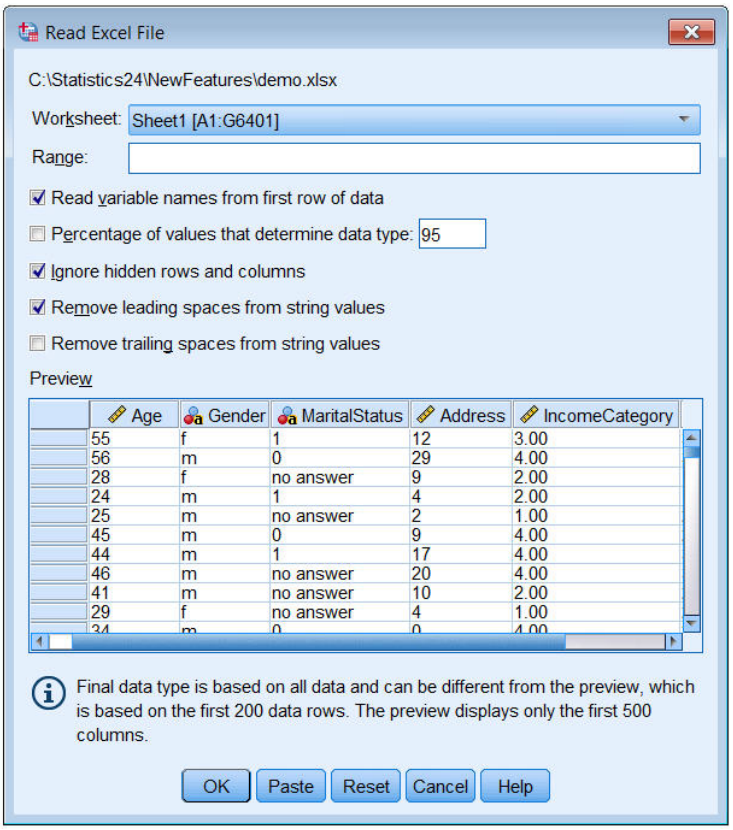


Figure 11. Read Excel File dialog

3. Make sure **Read variable names from the first row of data** is selected. If the column headings do not conform to variable name rules, they are converted to valid variable names. The original column headings are saved as variable labels.
4. Select **Remove leading spaces from string values**.
5. Deselect **Percentage of values that determine data type**.



The string value "no answer" is now displayed in the cells that were system-missing. If there is no percentage of values parameter and the column contains a mix of data type, the variable is read as a string data type. All values are preserved, but numeric values are treated as string values.

6. Select (check) **Percentage of values that determine data type** to treat *MaritalStatus* as a numeric variable.
7. Click **OK** to read the Excel file.

The data now appear in the Data Editor, with the column headings used as variable names. Since variable names can't contain spaces, the spaces from the original column headings are removed. For example, the column heading"Marital Status" is converted to the variable *MaritalStatus*. The original column heading is retained as a variable label.

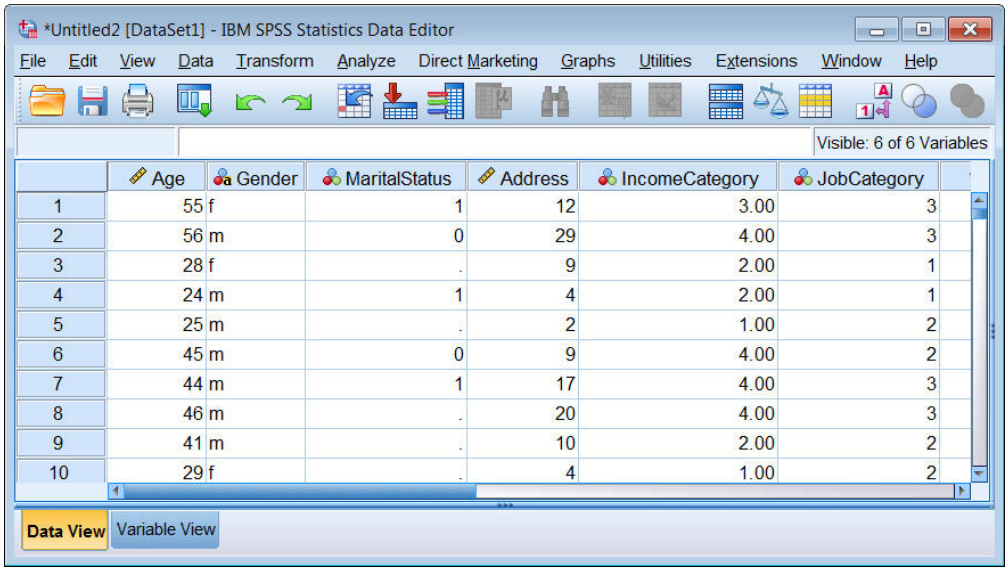


Figure 12. Imported Excel data

Related information:

Chapter 10, “Sample Files,” on page 83

Reading Data from a Database

Data from database sources are easily imported using the Database Wizard. Any database that uses ODBC (Open Database Connectivity) drivers can be read directly after the drivers are installed. ODBC drivers for many database formats are supplied on the installation CD. Additional drivers can be obtained from third-party vendors. One of the most common database applications, Microsoft Access, is discussed in this example.

Note: This example is specific to Microsoft Windows and requires an ODBC driver for Access. The Microsoft Access ODBC driver only works with the 32-bit version of IBM SPSS Statistics. The steps are similar on other platforms but may require a third-party ODBC driver for Access.

1. From the menus choose:

   File > **Import Data** > **Database** > **New Query...**

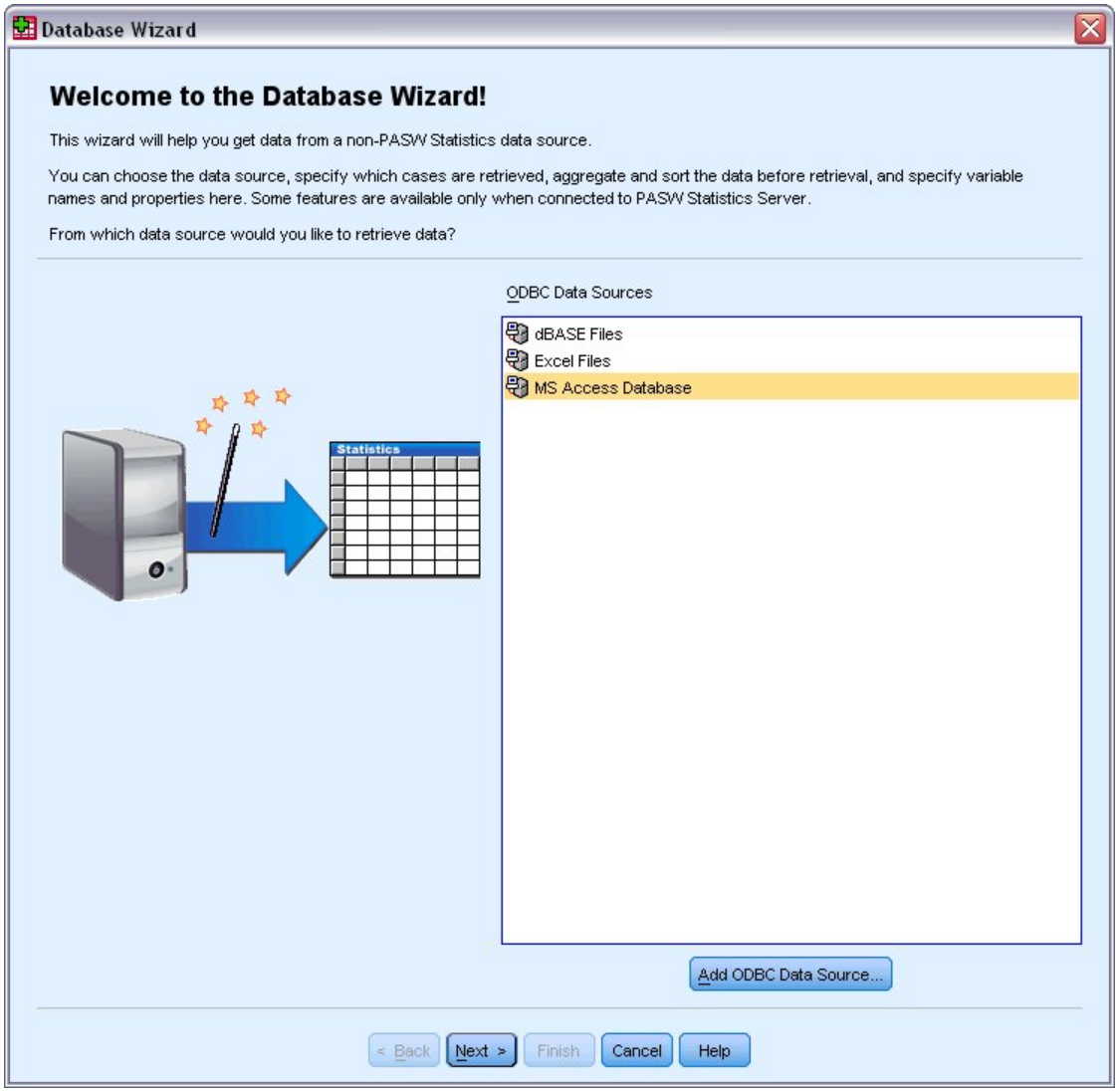


Figure 13. Database Wizard Welcome dialog box

2. Select **MS Access Database** from the list of data sources and click **Next**.

   Note: Depending on your installation, you may also see a list of OLEDB data sources on the left side of the wizard (Windows operating systems only), but this example uses the list of ODBC data sources displayed on the right side.
3. Click **Browse** to navigate to the Access database file that you want to open.
4. Open *demo.mdb*. See the topic Chapter 10, "Sample Files," on page 83 for more information.
5. Click **OK** in the login dialog box.

   In the next step, you can specify the tables and variables that you want to import.

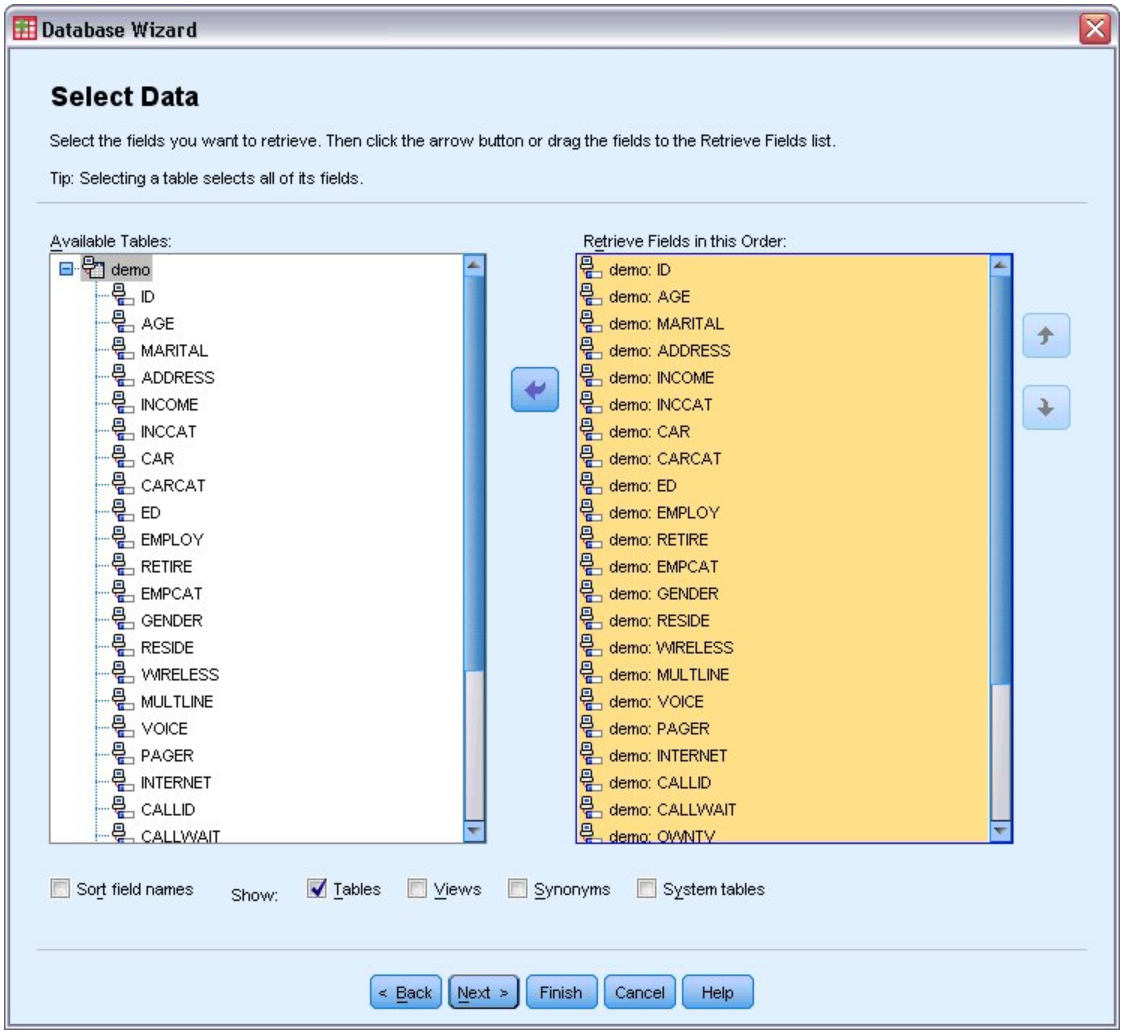


Figure 14. Select Data step

6. Drag the entire **demo** table to the Retrieve Fields In This Order list.
7. Click **Next**.

   In the next step, you can select which records (cases) to import.

   If you do not want to import all cases, you can import a subset of cases (for example, males older than 30), or you can import a random sample of cases from the data source. For large data sources, you may want to limit the number of cases to a small, representative sample to reduce the processing time.
8. Click **Next** to continue.

   Field names are used to create variable names. If necessary, the names are converted to valid variable names. The original field names are preserved as variable labels. You can also change the variable names before importing the database.

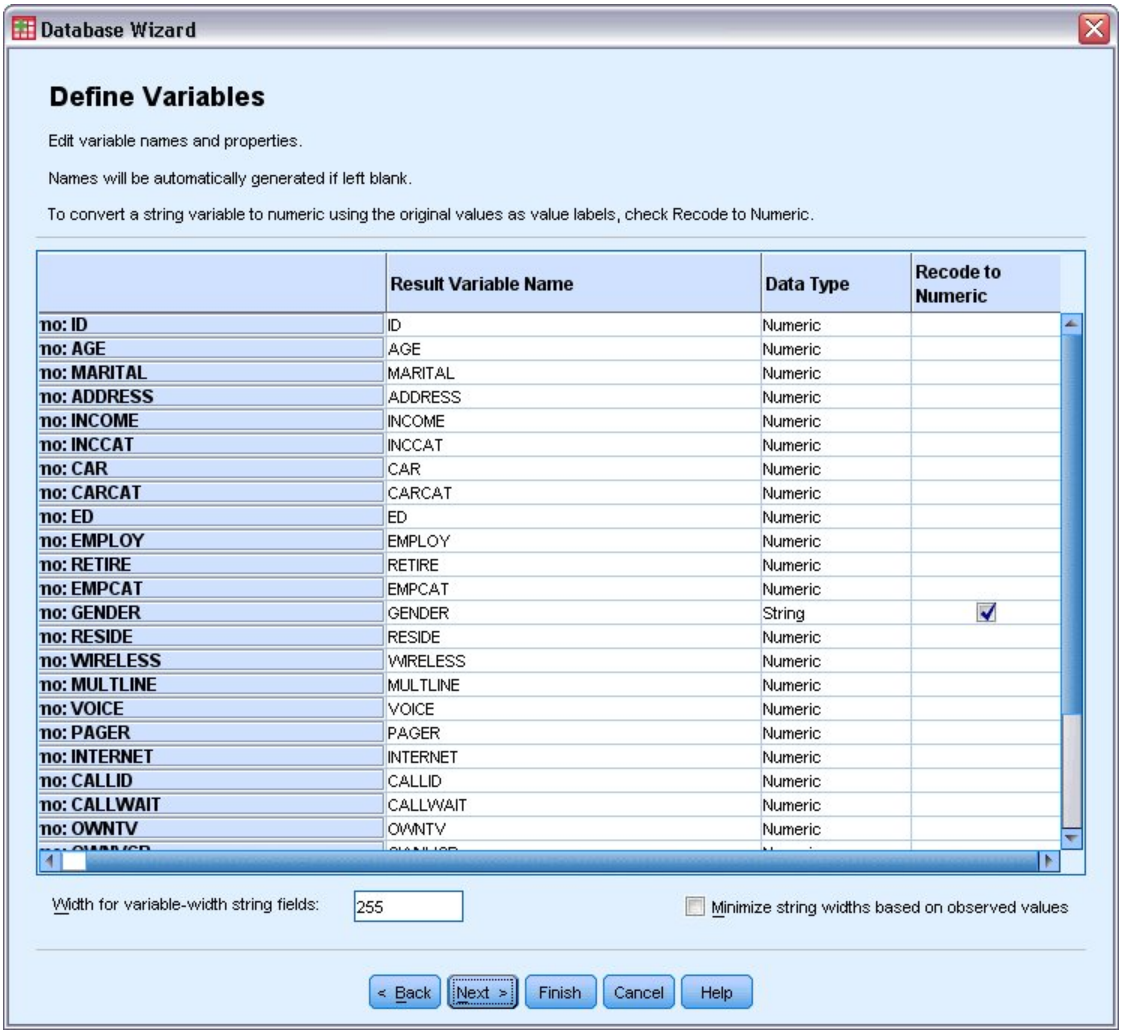


Figure 15. Define Variables step

9. Click the **Recode to Numeric** cell in the Gender field. This option converts string variables to integer variables and retains the original value as the value label for the new variable.
10. Click **Next** to continue.

    The SQL statement created from your selections in the Database Wizard appears in the Results step. This statement can be executed now or saved to a file for later use.
11. Click **Finish** to import the data.

All of the data in the Access database that you selected to import are now available in the Data Editor.

Reading Data from a Text File

Text files represent another common source of data. Many spreadsheet programs and databases can save their contents in one of many text file formats. Comma- or tab-delimited files refer to rows of data that use commas or tabs to indicate each variable. In this example, the data is tab-delimited.

1. From the menus choose:

   File > **Import Data** > **Text Data**
2. Go to the `Samples\English` folder and select `demo.txt`.

The Text Import Wizard guides you through the process of defining how the specified text file is interpreted.

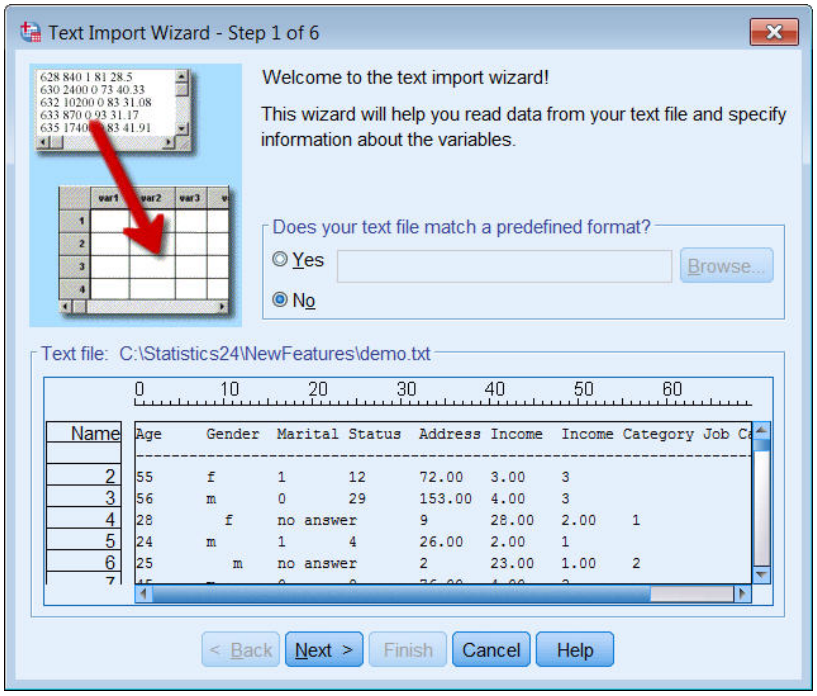


Figure 16. Text Import Wizard: Step 1 of 6

3. In Step 1, you can choose a predefined format or create a new format in the wizard. Select **No**.
4. Click **Next** to continue.

   As stated earlier, this file uses tab-delimited formatting. Also, the variable names are defined on the top line of this file.
5. In step 2 of the wizard, select **Delimited** to indicate that the file has a delimited formatting structure.
6. Select **Yes** to indicate that the file includes variable names at the top of the file.
7. Click **Next** to continue.
8. In step 3, enter 2 for the line number where the first case of data begins (because variable names are on the first line).
9. Keep the default values for the remainder of this step, and click **Next** to continue.

   The Data preview in Step 4 provides you with a quick way to ensure that the file is read correctly
10. Select **Tab** and deselect the other options for delimiters. **Space** is selected by default because the file contains spaces. For this file, spaces are part of the data values, not delimiters. You need to deselect **Space** to read the file correctly.
11. Select **Remove leading spaces for string values**. Spaces at the start of string values affect how string values are evaluated in expressions. In this file, some values for *Gender* have leading spaces that are not part of the value. If you do not remove those spaces, a value of " f" is treated as a different value than "f".

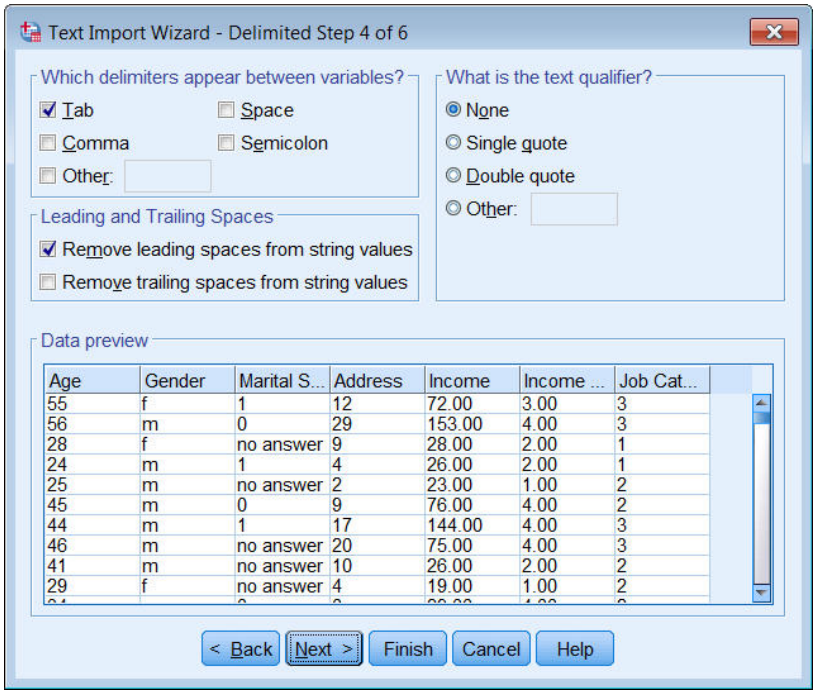


Figure 17. Text Import Wizard: Step 4 of 6

12. Click **Next** to continue.

    Because the variable names are modified to conform to naming rules, step 5 gives you the opportunity to edit any undesirable names.

    Data types can be defined here as well. For example, you can change *Income* to dollar currency format.

    To change a data type:

13. In the **Data preview**, select *Income*.
14. Select **Dollar** from the Data format drop-down list.

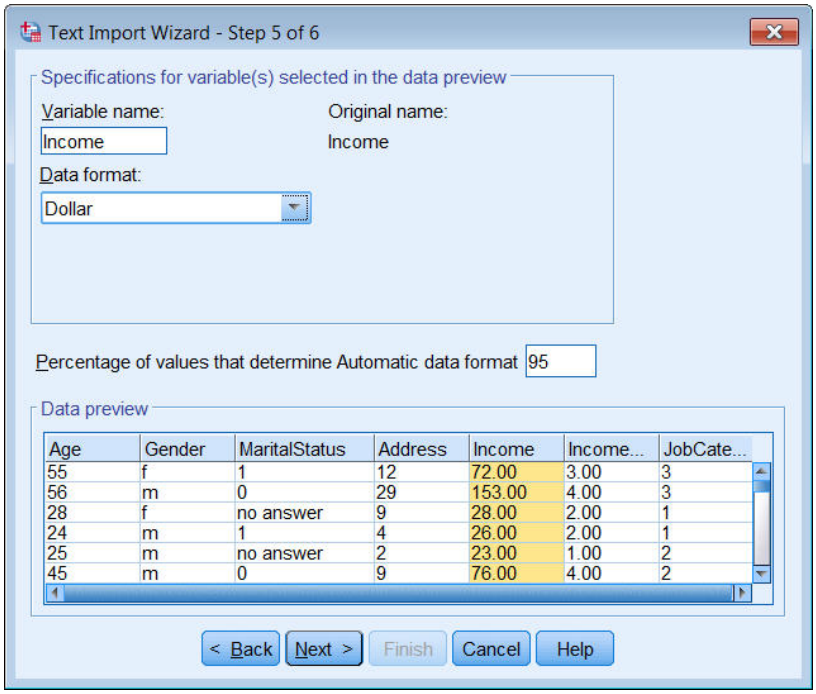


Figure 18. Change the data type

The variable *MaritalStatus* contains both string and numeric values. Less than five percent of the values are strings. With the default setting of 95% for **Percentage of values that determine the Automatic data format**, the variable is treated as numeric and the string values are set to system-missing. If no data format meets the percentage value, the variable is treated as a string variable. If you change the setting to 100, all values are preserved, but all numeric values are treated as strings.

15. Click **Next** to continue.
16. Leave the default selections in the last step, and click **Finish** to import the data.

Chapter 3. Using the Data Editor

The Data Editor displays the contents of the active data file. The information in the Data Editor consists of variables and cases.

- In Data View, columns represent variables, and rows represent cases (observations).
- In Variable View, each row is a variable, and each column is an attribute that is associated with that variable.

Variables are used to represent the different types of data that you have compiled. A common analogy is that of a survey. The response to each question on a survey is equivalent to a variable. Variables come in many different types, including numbers, strings, currency, and dates.

Entering Numeric Data

Data can be entered into the Data Editor, which may be useful for small data files or for making minor edits to larger data files.

1. Click the **Variable View** tab at the bottom of the Data Editor window.

   You need to define the variables that will be used. In this case, only three variables are needed: *age*, *marital status*, and *income*.

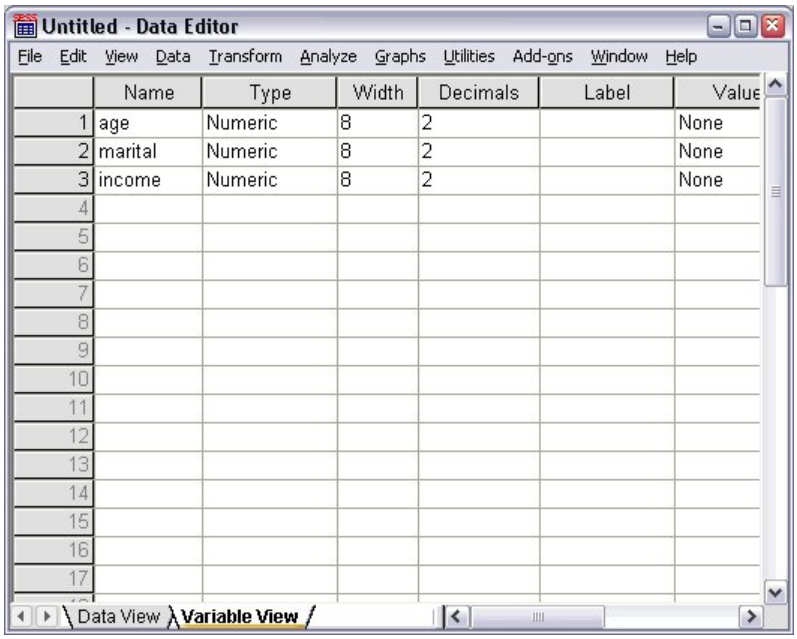


Figure 19. Variable names in Variable View

2. In the first row of the first column, type `age`.
3. In the second row, type `marital`.
4. In the third row, type `income`.

   New variables are automatically given a Numeric data type.

   If you don't enter variable names, unique names are automatically created. However, these names are not descriptive and are not recommended for large data files.
5. Click the **Data View** tab to continue entering the data.

The names that you entered in Variable View are now the headings for the first three columns in Data View.

Begin entering data in the first row, starting at the first column.

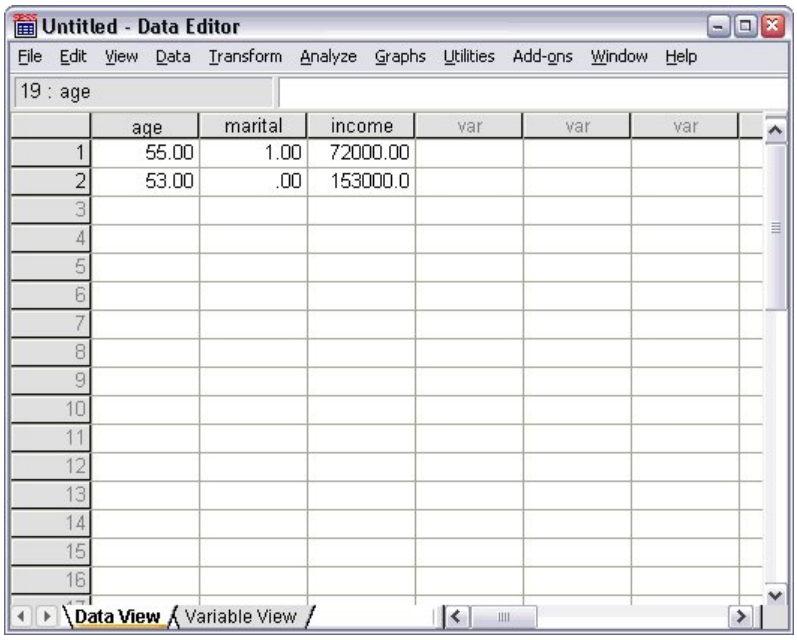


Figure 20. Values entered in Data View

6. In the *age* column, type `55`.
7. In the *marital* column, type `1`.
8. In the *income* column, type `72000`.
9. Move the cursor to the second row of the first column to add the next subject's data.
10. In the *age* column, type `53`.
11. In the *marital* column, type `0`.
12. In the *income* column, type `153000`.

    Currently, the *age* and *marital* columns display decimal points, even though their values are intended to be integers. To hide the decimal points in these variables:
13. Click the **Variable View** tab at the bottom of the Data Editor window.
14. In the *Decimals* column of the *age* row, type `0` to hide the decimal.
15. In the *Decimals* column of the *marital* row, type `0` to hide the decimal.

Entering String Data

Non-numeric data, such as strings of text, can also be entered into the Data Editor.

1. Click the **Variable View** tab at the bottom of the Data Editor window.
2. In the first cell of the first empty row, type `sex` for the variable name.
3. Click the *Type* cell next to your entry.
4. Click the button on the right side of the *Type* cell to open the Variable Type dialog box.
5. Select **String** to specify the variable type.
6. Click **OK** to save your selection and return to the Data Editor.

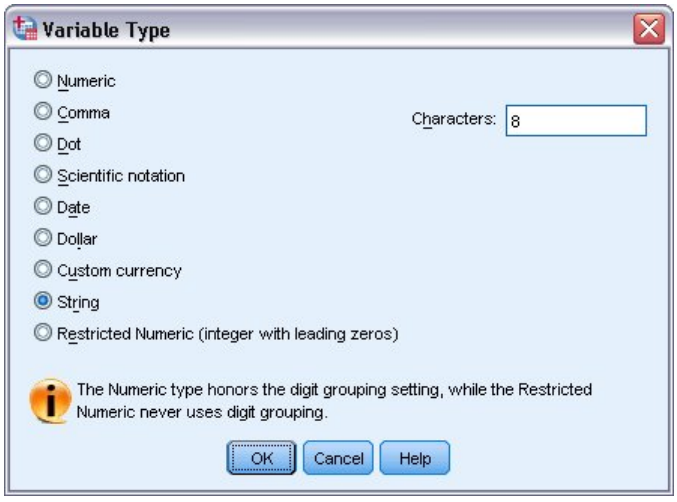


Figure 21. Variable Type dialog box

Defining Data

In addition to defining data types, you can also define descriptive variable labels and value labels for variable names and data values. These descriptive labels are used in statistical reports and charts.

Adding Variable Labels

Labels are meant to provide descriptions of variables. These descriptions are often longer versions of variable names. Labels can be up to 255 bytes. These labels are used in your output to identify the different variables.

1. Click the **Variable View** tab at the bottom of the Data Editor window.
2. In the *Label* column of the *age* row, type `Respondent's Age`.
3. In the *Label* column of the *marital* row, type `Marital Status`.
4. In the *Label* column of the *income* row, type `Household Income`.
5. In the *Label* column of the *sex* row, type `Gender`.

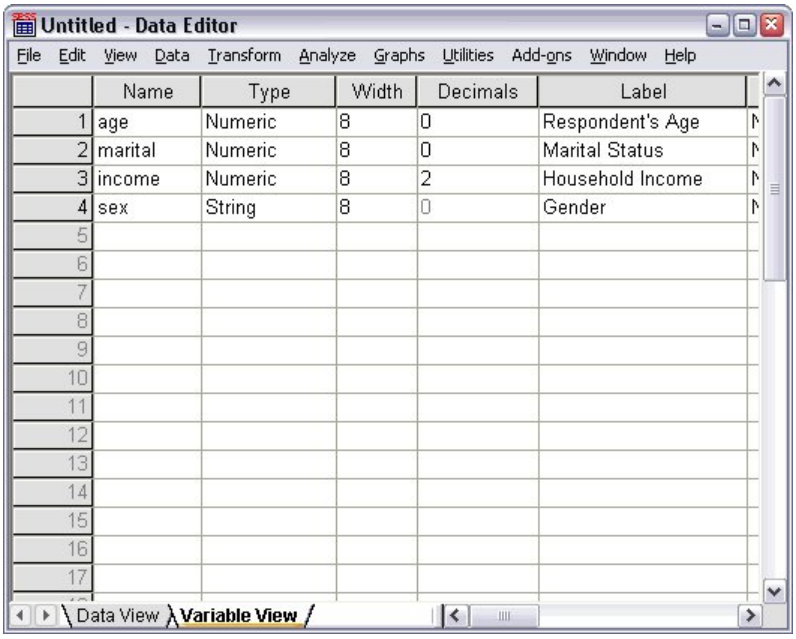


Figure 22. Variable labels entered in Variable View

Changing Variable Type and Format

The *Type* column displays the current data type for each variable. The most common data types are numeric and string, but many other formats are supported. In the current data file, the *income* variable is defined as a numeric type.

1. Click the *Type* cell for the *income* row, and then click the button on the right side of the cell to open the Variable Type dialog box.
2. Select **Dollar**.

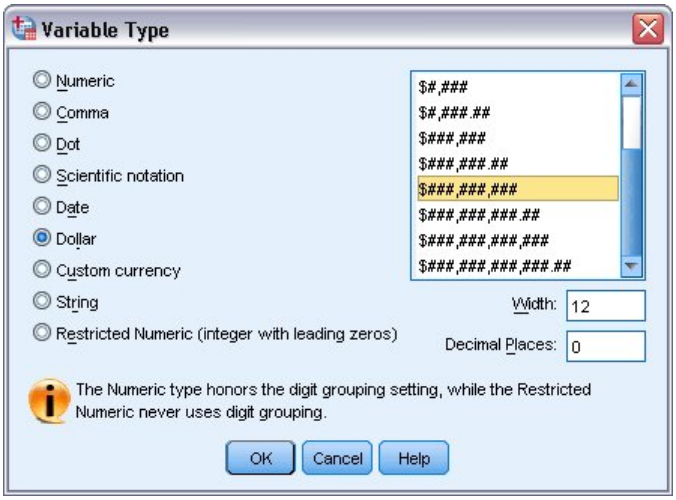


Figure 23. Variable Type dialog box

The formatting options for the currently selected data type are displayed.

3. For the format of the currency in this example, select **$###,###,###**.
4. Click **OK** to save your changes.

Adding Value Labels

Value labels provide a method for mapping your variable values to a string label. In this example, there are two acceptable values for the *marital* variable. A value of 0 means that the subject is single, and a value of 1 means that he or she is married.

1. Click the *Values* cell for the *marital* row, and then click the button on the right side of the cell to open the Value Labels dialog box.

   The **value** is the actual numeric value.

   The **value label** is the string label that is applied to the specified numeric value.
2. Type `0` in the Value field.
3. Type `Single` in the Label field.
4. Click **Add** to add this label to the list.

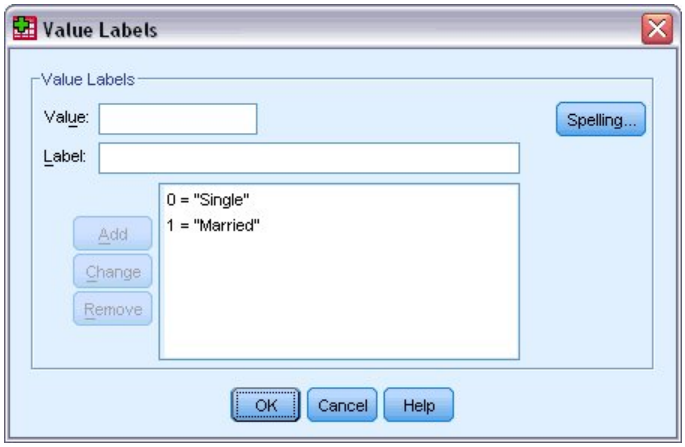


Figure 24. Value Labels dialog box

5. Type `1` in the Value field, and type `Married` in the Label field.
6. Click **Add**, and then click **OK** to save your changes and return to the Data Editor.

   These labels can also be displayed in Data View, which can make your data more readable.
7. Click the **Data View** tab at the bottom of the Data Editor window.
8. From the menus choose:

View > **Value Labels**

The labels are now displayed in a list when you enter values in the Data Editor. This setup has the benefit of suggesting a valid response and providing a more descriptive answer.

If the Value Labels menu item is already active (with a check mark next to it), choosing **Value Labels** again will turn *off* the display of value labels.

Handling Missing Data

Missing or invalid data are generally too common to ignore. Survey respondents may refuse to answer certain questions, may not know the answer, or may answer in an unexpected format. If you don't filter or identify these data, your analysis may not provide accurate results.

For numeric data, empty data fields or fields containing invalid entries are converted to system-missing, which is identifiable by a single period.

The reason a value is missing may be important to your analysis. For example, you may find it useful to distinguish between those respondents who refused to answer a question and those respondents who didn't answer a question because it was not applicable.

Missing Values for a Numeric Variable

1. Click the **Variable View** tab at the bottom of the Data Editor window.
2. Click the *Missing* cell in the *age* row, and then click the button on the right side of the cell to open the Missing Values dialog box.

   In this dialog box, you can specify up to three distinct missing values, or you can specify a range of values plus one additional discrete value.

Figure 25. Missing Values dialog box

3. Select **Discrete missing values**.
4. Type `999` in the first text box and leave the other two text boxes empty.
5. Click **OK** to save your changes and return to the Data Editor.

   Now that the missing data value has been added, a label can be applied to that value.
6. Click the *Values* cell in the *age* row, and then click the button on the right side of the cell to open the Value Labels dialog box.
7. Type `999` in the Value field.
8. Type `No Response` in the Label field.
9. Click **Add** to add this label to your data file.
10. Click **OK** to save your changes and return to the Data Editor.

Missing Values for a String Variable

Missing values for string variables are handled similarly to the missing values for numeric variables. However, unlike numeric variables, empty fields in string variables are not designated as system-missing. Rather, they are interpreted as an empty string.

1. Click the **Variable View** tab at the bottom of the Data Editor window.
2. Click the *Missing* cell in the *sex* row, and then click the button on the right side of the cell to open the Missing Values dialog box.
3. Select **Discrete missing values**.
4. Type `NR` in the first text box.

   Missing values for string variables are case sensitive. So, a value of *nr* is not treated as a missing value.
5. Click **OK** to save your changes and return to the Data Editor.

   Now you can add a label for the missing value.
6. Click the *Values* cell in the *sex* row, and then click the button on the right side of the cell to open the Value Labels dialog box.

7. Type `NR` in the Value field.
8. Type `No Response` in the Label field.
9. Click **Add** to add this label to your project.
10. Click **OK** to save your changes and return to the Data Editor.

Chapter 4. Examining Summary Statistics for Individual Variables

This section discusses simple summary measures and how the level of measurement of a variable influences the types of statistics that should be used. We will use the data file *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.

Level of Measurement

Different summary measures are appropriate for different types of data, depending on the level of measurement:

Categorical. Data with a limited number of distinct values or categories (for example, gender or marital status). Also referred to as **qualitative data**. Categorical variables can be string (alphanumeric) data or numeric variables that use numeric codes to represent categories (for example, 0 = *Unmarried* and 1 = *Married*). There are two basic types of categorical data:

- **Nominal.** Categorical data where there is no inherent order to the categories. For example, a job category of *sales* is not higher or lower than a job category of *marketing* or *research*.
- **Ordinal.** Categorical data where there is a meaningful order of categories, but there is not a measurable distance between categories. For example, there is an order to the values *high*, *medium*, and *low*, but the "distance" between the values cannot be calculated.

Scale. Data measured on an interval or ratio scale, where the data values indicate both the order of values and the distance between values. For example, a salary of $72,195 is higher than a salary of $52,398, and the distance between the two values is $19,797. Also referred to as **quantitative** or **continuous data**.

Summary Measures for Categorical Data

For categorical data, the most typical summary measure is the number or percentage of cases in each category. The **mode** is the category with the greatest number of cases. For ordinal data, the **median** (the value at which half of the cases fall above and below) may also be a useful summary measure if there is a large number of categories.

The Frequencies procedure produces frequency tables that display both the number and percentage of cases for each observed value of a variable.

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Frequencies...**

   Note: This feature requires Statistics Base Edition.
2. Select *Owns PDA [ownpda]* and *Owns TV [owntv]* and move them into the Variable(s) list.

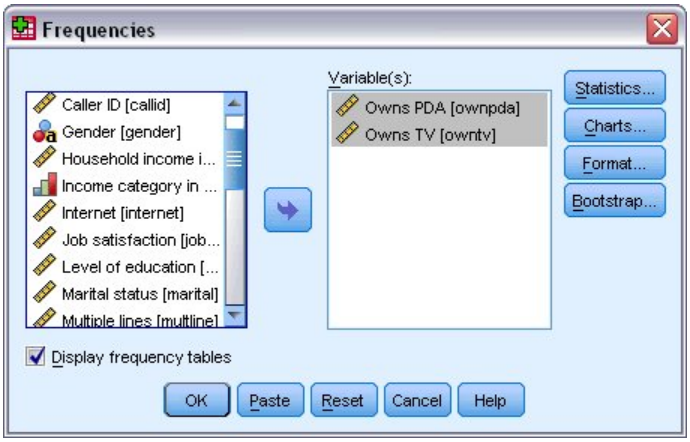


Figure 26. Categorical variables selected for analysis

3. Click **OK** to run the procedure.

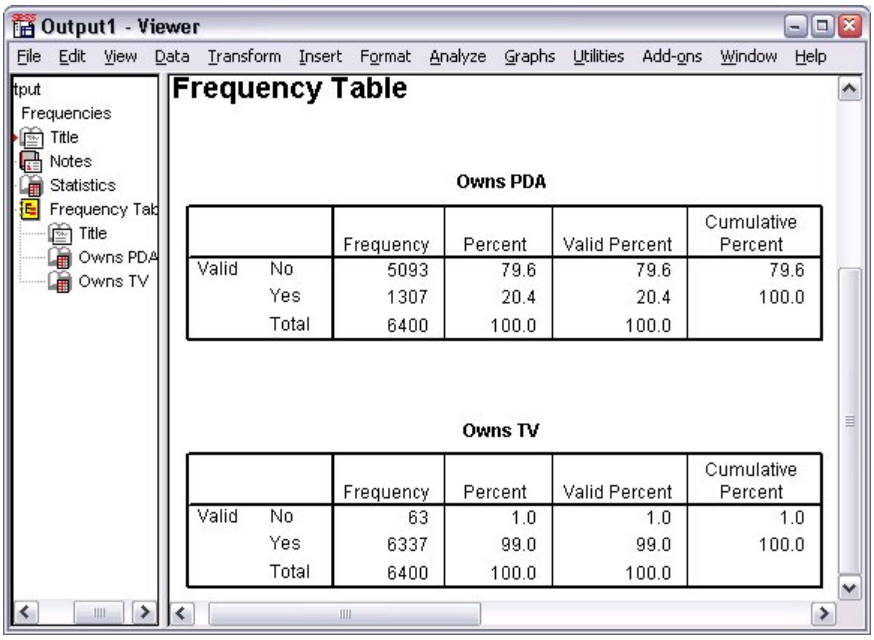


Figure 27. Frequency tables

The frequency tables are displayed in the Viewer window. The frequency tables reveal that only 20.4% of the people own PDAs, but almost everybody owns a TV (99.0%). These might not be interesting revelations, although it might be interesting to find out more about the small group of people who do not own televisions.

Charts for Categorical Data

You can graphically display the information in a frequency table with a bar chart or pie chart.

1. Open the Frequencies dialog box again. (The two variables should still be selected.)

   You can use the Dialog Recall button on the toolbar to quickly return to recently used procedures.

Figure 28. Dialog Recall button

2. Click **Charts**.
3. Select **Bar charts** and then click **Continue**.
4. Click **OK** in the main dialog box to run the procedure.

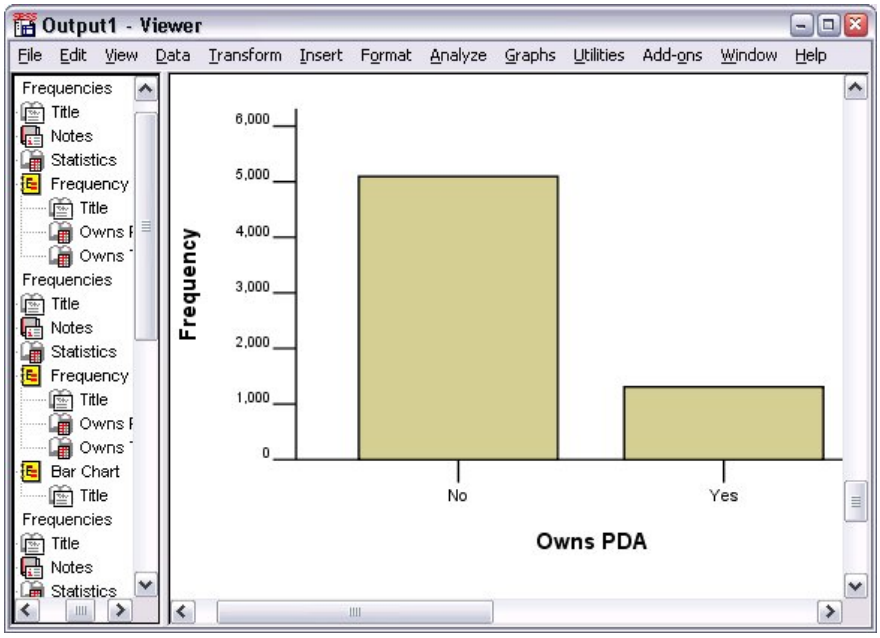


Figure 29. Bar chart

In addition to the frequency tables, the same information is now displayed in the form of bar charts, making it easy to see that most people do not own PDAs but almost everyone owns a TV.

Summary Measures for Scale Variables

There are many summary measures available for scale variables, including:

- **Measures of central tendency.** The most common measures of central tendency are the **mean** (arithmetic average) and **median** (value at which half the cases fall above and below).
- **Measures of dispersion.** Statistics that measure the amount of variation or spread in the data include the standard deviation, minimum, and maximum.

1. Open the Frequencies dialog box again.
2. Click **Reset** to clear any previous settings.
3. Select *Household income in thousands [income]* and move it into the Variable(s) list.
4. Click **Statistics**.
5. Select **Mean**, **Median**, **Std. deviation**, **Minimum**, and **Maximum**.
6. Click **Continue**.
7. Deselect **Display frequency tables** in the main dialog box. (Frequency tables are usually not useful for scale variables since there may be almost as many distinct values as there are cases in the data file.)

8. Click **OK** to run the procedure.

The Frequencies Statistics table is displayed in the Viewer window.

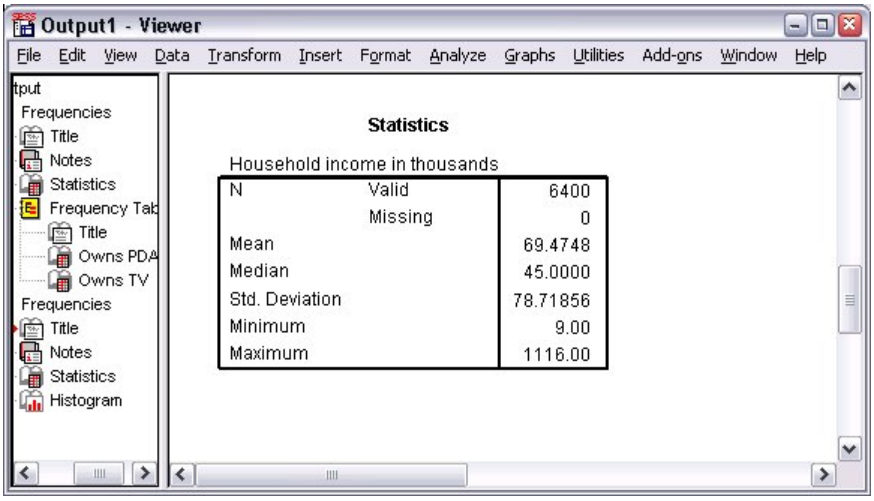


Figure 30. Frequencies Statistics table

In this example, there is a large difference between the mean and the median. The mean is almost 25,000 greater than the median, indicating that the values are not normally distributed. You can visually check the distribution with a histogram.

Histograms for Scale Variables

1. Open the Frequencies dialog box again.
2. Click **Charts**.
3. Select **Histograms** and **With normal curve**.
4. Click **Continue**, and then click **OK** in the main dialog box to run the procedure.

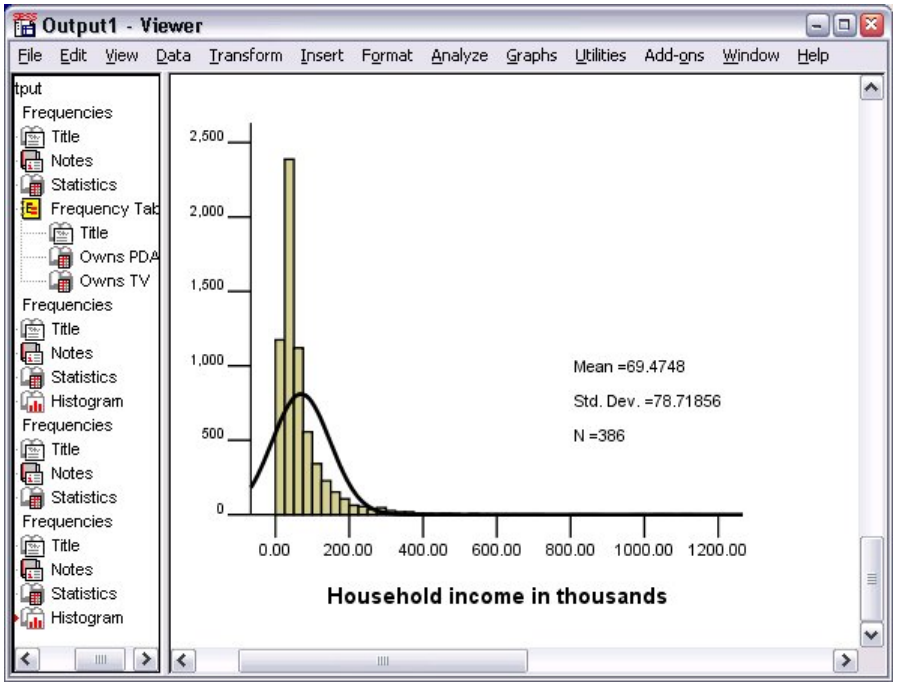


Figure 31. Histogram

The majority of cases are clustered at the lower end of the scale, with most falling below 100,000. There are, however, a few cases in the 500,000 range and beyond (too few to even be visible without modifying the histogram). These high values for only a few cases have a significant effect on the mean but little or no effect on the median, making the median a better indicator of central tendency in this example.

Chapter 5. Creating and editing charts

You can create and edit a wide variety of chart types. In this chapter, we will create and edit bar charts. You can apply the principles to any chart type.

Chart creation basics

To demonstrate the basics of chart creation, we will create a bar chart of mean income for different levels of job satisfaction. This example uses the data file *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.

1. From the menus choose:

Graphs > **Chart Builder...**

The Chart Builder dialog box is an interactive window that allows you to preview how a chart will look while you build it.

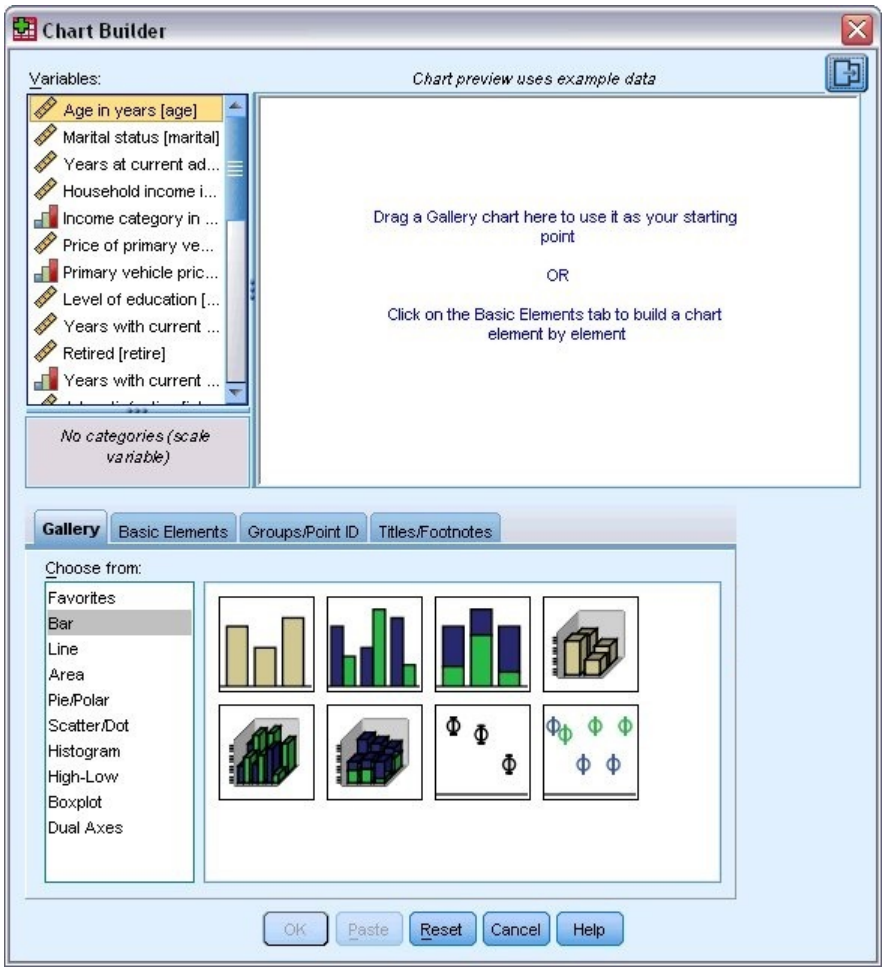


Figure 32. Chart Builder dialog box

Using the Chart Builder gallery

1. Click the **Gallery** tab if it is not selected.

The Gallery includes many different predefined charts, which are organized by chart type. The Basic Elements tab also provides basic elements (such as axes and graphic elements) for creating charts from scratch, but it's easier to use the Gallery.

2. Click **Bar** if it is not selected.

   Icons representing the available bar charts in the Gallery appear in the dialog box. The pictures should provide enough information to identify the specific chart type. If you need more information, you can also display a ToolTip description of the chart by pausing your cursor over an icon.

3. Drag the icon for the simple bar chart onto the "canvas," which is the large area above the Gallery. The Chart Builder displays a preview of the chart on the canvas. Note that the data used to draw the chart are not your actual data. They are example data.

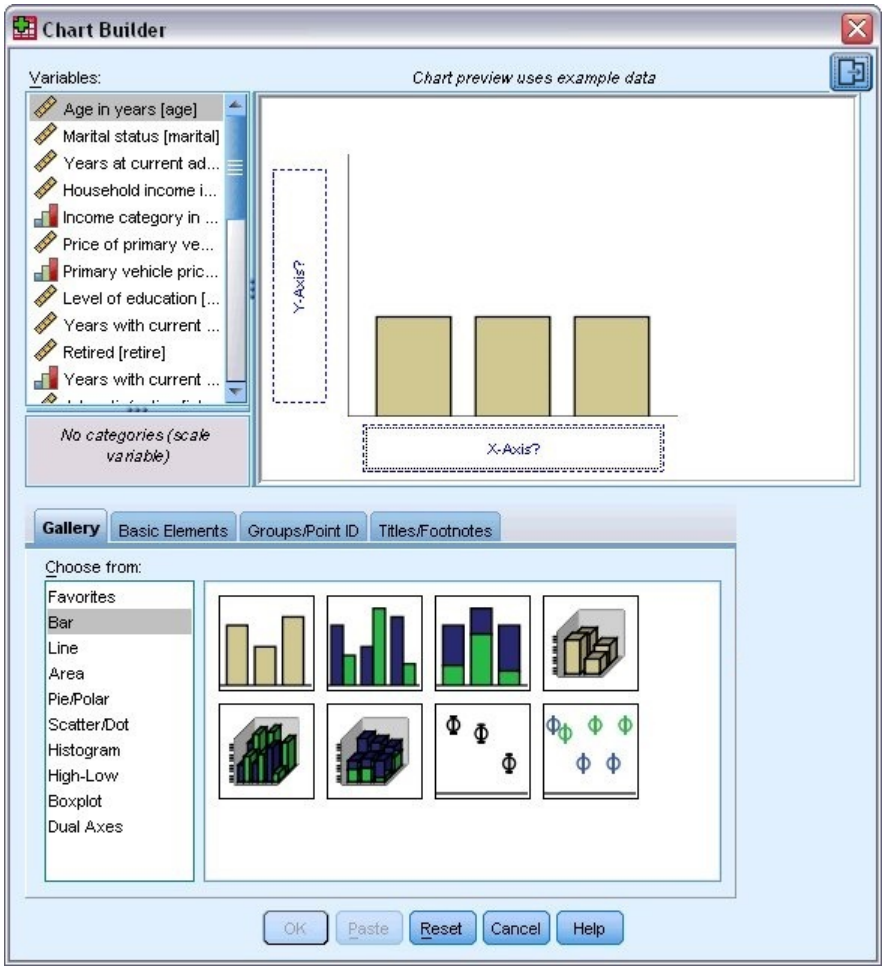


Figure 33. Bar chart on Chart Builder canvas

Defining variables and statistics

Although there is a chart on the canvas, it is not complete because there are no variables or statistics to control how tall the bars are and to specify which variable category corresponds to each bar. You can't have a chart without variables and statistics. You add variables by dragging them from the Variables list, which is located to the left of the canvas.

A variable's measurement level is important in the Chart Builder. You are going to use the *Job satisfaction* variable on the *x* axis. However, the icon (which looks like a ruler) next to the variable indicates that its measurement level is defined as scale. To create the correct chart, you must use a categorical

measurement level. Instead of going back and changing the measurement level in the Variable View, you can change the measurement level temporarily in the Chart Builder.

1. Right-click *Job satisfaction* in the Variables list and choose **Ordinal**. Ordinal is an appropriate measurement level because the categories in *Job satisfaction* can be ranked by level of satisfaction. Note that the icon changes after you change the measurement level.
2. Now drag *Job satisfaction* from the Variables list to the *x* axis drop zone.

   The *y* axis drop zone defaults to the *Count* statistic. If you want to use another statistic (such as percentage or mean), you can easily change it. You will not use either of these statistics in this example, but we will review the process in case you need to change this statistic at another time.
3. Click the **Element Properties** tab in the side bar of the Chart Builder. (If the side bar is not displayed, click the button in the upper right corner of the Chart Builder to display the side bar.)

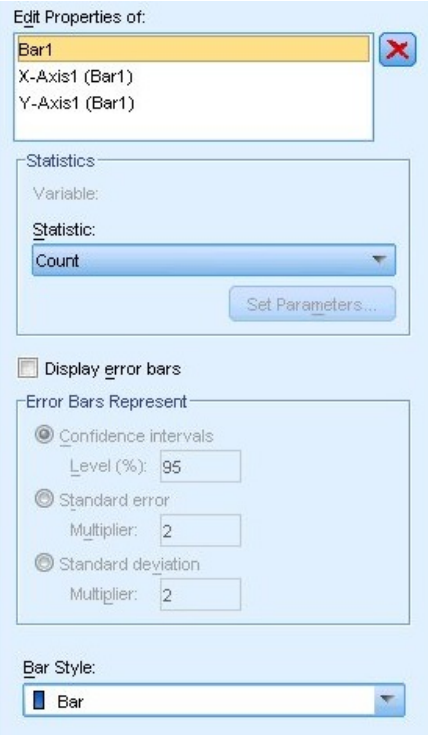


Figure 34. Element Properties

Element Properties allows you to change the properties of the various chart elements. These elements include the graphic elements (such as the bars in the bar chart) and the axes on the chart. Select one of the elements in the Edit Properties of list to change the properties associated with that element. Also note the red *X* located to the right of the list. This button deletes a graphic element from the canvas. Because **Bar1** is selected, the properties shown apply to graphic elements, specifically the bar graphic element.

The Statistic drop-down list shows the specific statistics that are available. The same statistics are usually available for every chart type. Be aware that some statistics require that the *y* axis drop zone contains a variable.

4. Drag *Household income in thousands* from the Variables list to the *y* axis drop zone. Because the variable on the *y* axis is scalar and the *x* axis variable is categorical (ordinal is a type of categorical measurement level), the *y* axis drop zone defaults to the *Mean* statistic. These are the variables and statistics you want, so there is no need to change the element properties.

Adding text

You can also add titles and footnotes to the chart.

1. Click the **Titles/Footnotes** tab.
2. Select **Title 1**.

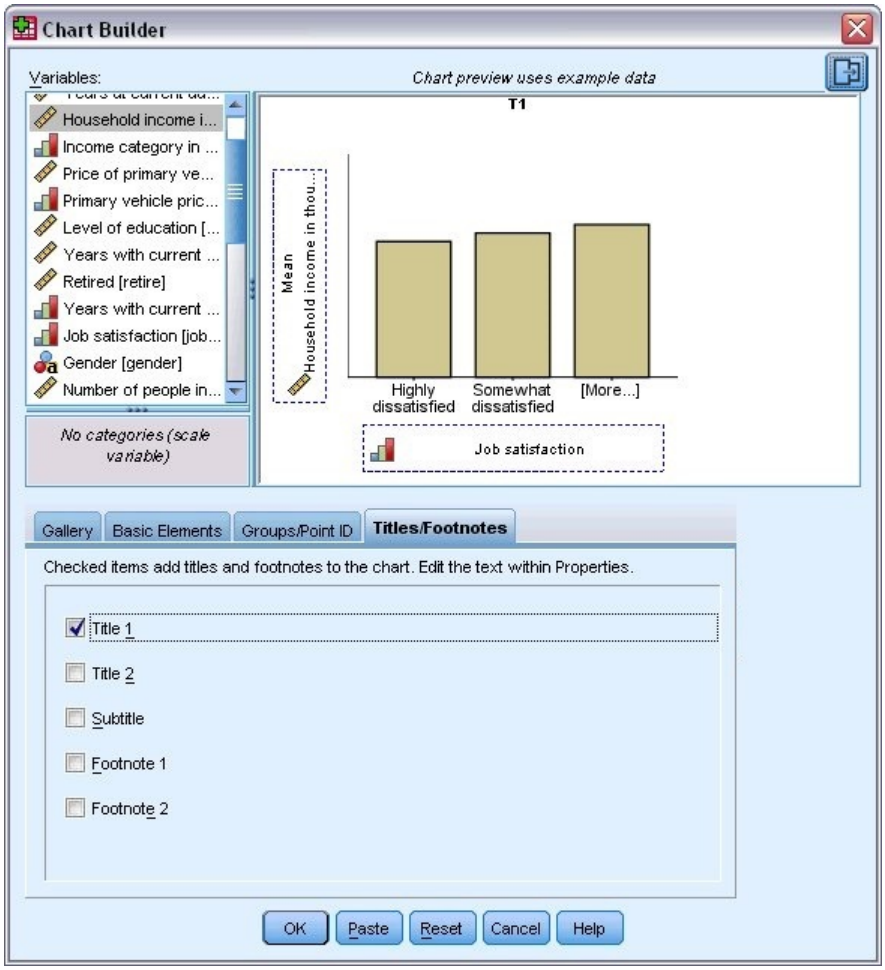


Figure 35. Title 1 displayed on canvas

The title appears on the canvas with the label **T1**.

3. In the **Element Properties** tab, select **Title 1** in the Edit Properties of list.
4. In the Content text box, type `Income by Job Satisfaction`. This is the text that the title will display.

Creating the chart

1. Click **OK** to create the bar chart.

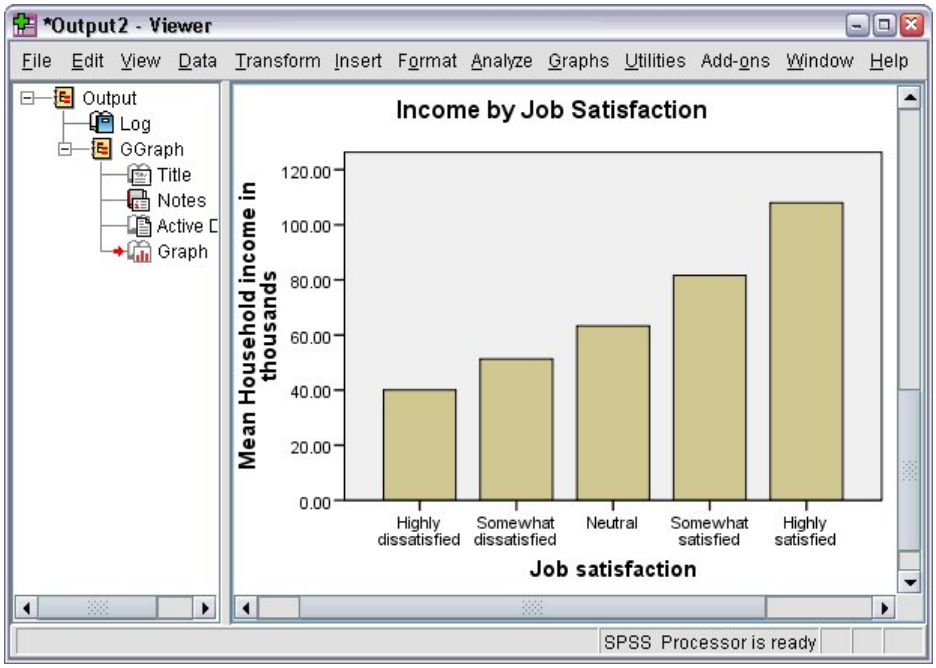


Figure 36. Bar chart

The bar chart reveals that respondents who are more satisfied with their jobs tend to have higher household incomes.

Chapter 6. Working with Output

The results from running a statistical procedure are displayed in the Viewer. The output produced can be statistical tables, charts, graphs, or text, depending on the choices you make when you run the procedure. This section uses the files *viewertut.spv* and *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.

Using the Viewer

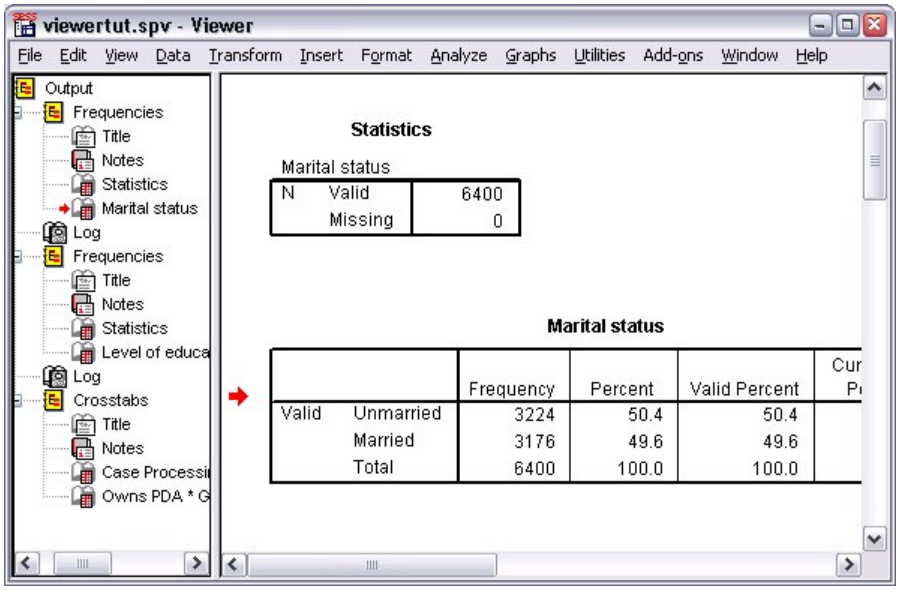


Figure 37. Viewer

The Viewer window is divided into two panes. The **outline pane** contains an outline of all of the information stored in the Viewer. The **contents pane** contains statistical tables, charts, and text output.

Use the scroll bars to navigate through the window's contents, both vertically and horizontally. For easier navigation, click an item in the outline pane to display it in the contents pane.

1. Click and drag the right border of the outline pane to change its width.

   An open book icon in the outline pane indicates that it is currently visible in the Viewer, although it may not currently be in the visible portion of the contents pane.
2. To hide a table or chart, double-click its book icon in the outline pane.

   The open book icon changes to a closed book icon, signifying that the information associated with it is now hidden.
3. To redisplay the hidden output, double-click the closed book icon.

   You can also hide all of the output from a particular statistical procedure or all of the output in the Viewer.
4. Click the box with the minus sign (–) to the left of the procedure whose results you want to hide, or click the box next to the topmost item in the outline pane to hide all of the output.

   The outline collapses, visually indicating that these results are hidden.

   You can also change the order in which the output is displayed.
5. In the outline pane, click the items that you want to move.

6. Drag the selected items to a new location in the outline.

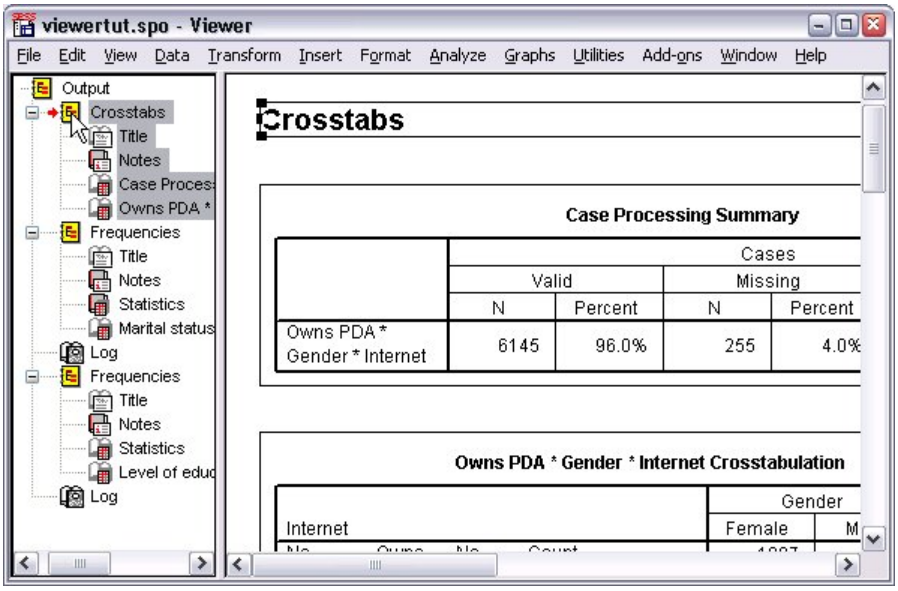


Figure 38. Reordered output in the Viewer

You can also move output items by clicking and dragging them in the contents pane.

Using the Pivot Table Editor

The results from most statistical procedures are displayed in **pivot tables**.

Accessing Output Definitions

Many statistical terms are displayed in the output. Definitions of these terms can be accessed directly in the Viewer.

1. Double-click the *Owns PDA * Gender * Internet Crosstabulation* table.
2. Right-click *Expected Count* and choose **What's This?** from the pop-up menu.

The definition is displayed in a pop-up window.

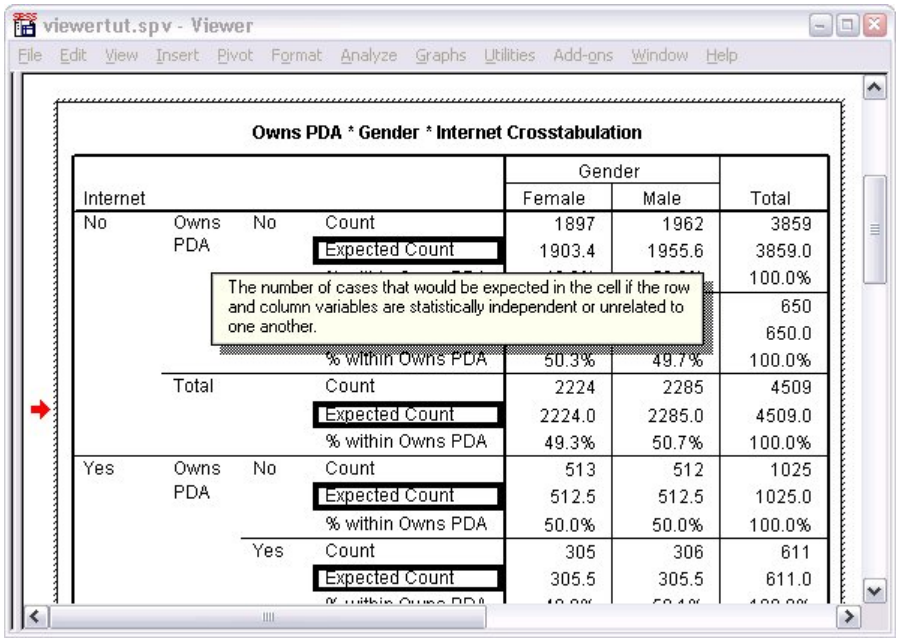


Figure 39. Pop-up definition

Pivoting Tables

The default tables produced may not display information as neatly or as clearly as you would like. With pivot tables, you can transpose rows and columns ("flip" the table), adjust the order of data in a table, and modify the table in many other ways. For example, you can change a short, wide table into a long, thin one by transposing rows and columns. Changing the layout of the table does not affect the results. Instead, it's a way to display your information in a different or more desirable manner.

1. If it's not already activated, double-click the *Owns PDA * Gender * Internet Crosstabulation* table to activate it.
2. If the Pivoting Trays window is not visible, from the menus choose:

   Pivot > Pivoting Trays

   Pivoting trays provide a way to move data between columns, rows, and layers.

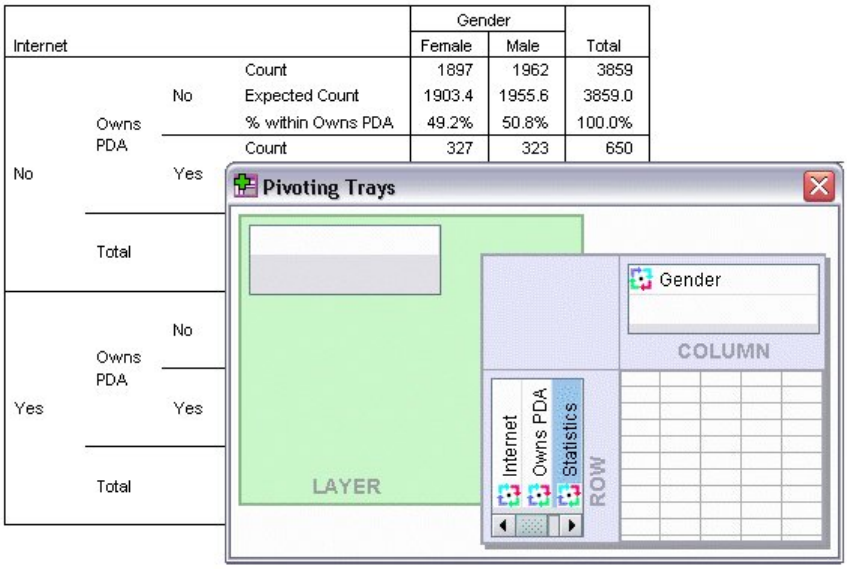


Figure 40. Pivoting trays

3. Drag the *Statistics* element from the Row dimension to the Column dimension, below *Gender*. The table is immediately reconfigured to reflect your changes.

   The order of the elements in the pivoting tray reflects the order of the elements in the table.
4. Drag and drop the *Owns PDA* element before the *Internet* element in the row dimension to reverse the order of these two rows.

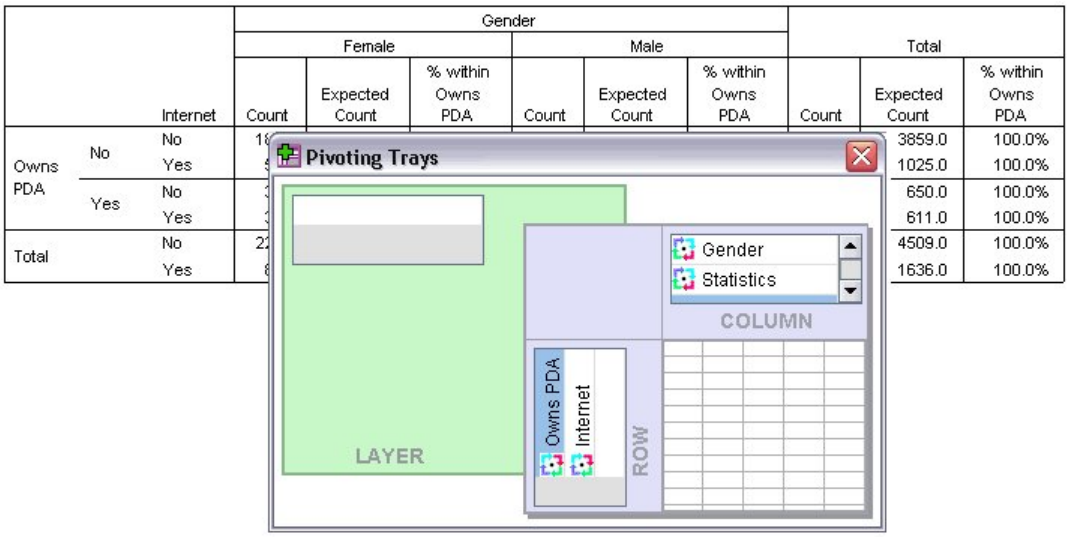


Figure 41. Swap rows

Creating and Displaying Layers

Layers can be useful for large tables with nested categories of information. By creating layers, you simplify the look of the table, making it easier to read.

1. Drag the *Gender* element from the Column dimension to the Layer dimension.

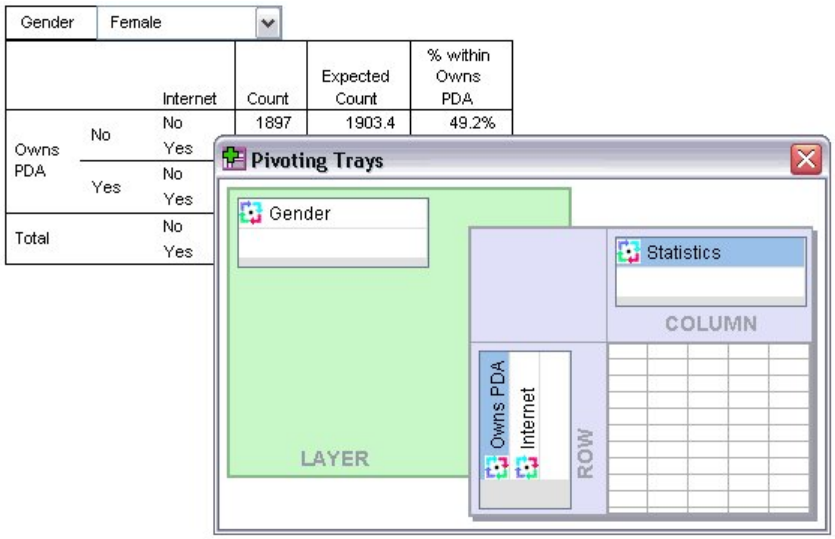


Figure 42. Gender pivot icon in the Layer dimension

To display a different layer, select a category from the drop-down list in the table.

Editing Tables

Unless you've taken the time to create a custom TableLook, pivot tables are created with standard formatting. You can change the formatting of any text within a table. Formats that you can change include font name, font size, font style (bold or italic), and color.

1. Double-click the *Level of education* table.
2. If the Formatting toolbar is not visible, from the menus choose:

   View > **Toolbar**
3. Click the title text, *Level of education*.
4. From the drop-down list of font sizes on the toolbar, choose **12**.
5. To change the color of the title text, click the text color tool and choose a new color.

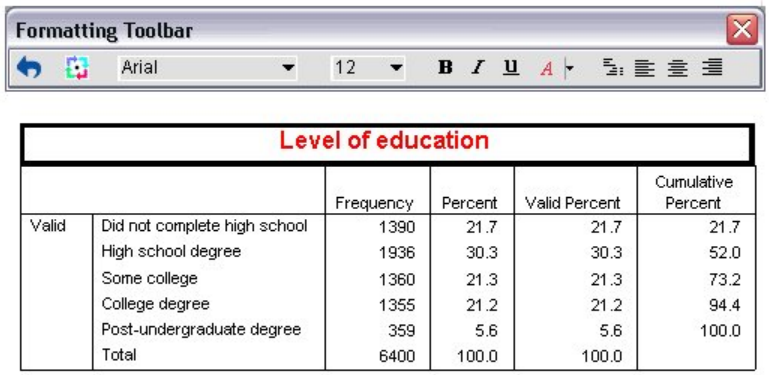


Figure 43. Reformatted title text in the pivot table

You can also edit the contents of tables and labels. For example, you can change the title of this table.

6. Double-click the title.
7. Type `Education Level` for the new label.

Note: If you change the values in a table, totals and other statistics are not recalculated.

Hiding Rows and Columns

Some of the data displayed in a table may not be useful or it may unnecessarily complicate the table. Fortunately, you can hide entire rows and columns without losing any data.

1. If it's not already activated, double-click the *Education Level* table to activate it.
2. Click *Valid Percent* column label to select it.
3. From the Edit menu or the right-click pop-up menu choose:
   Select > **Data and Label Cells**
4. From the View menu choose **Hide** or from the right-click pop-up menu choose **Hide Category**.
   The column is now hidden but not deleted.

Education Level

		Frequency	Percent	Cumulative Percent
Valid	Did not complete high school	1390	21.7	21.7
	High school degree	1936	30.3	52.0
	Some college	1360	21.3	73.2
	College degree	1355	21.2	94.4
	Post-undergraduate degree	359	5.6	100.0
	Total	6400	100.0	

Figure 44. Valid Percent column hidden in table

To redisplay the column:

5. From the menus choose:

View > **Show All**

Rows can be hidden and displayed in the same way as columns.

Changing Data Display Formats

You can easily change the display format of data in pivot tables.

1. If it's not already activated, double-click the *Education Level* table to activate it.
2. Click the *Percent* column label to select it.
3. From the Edit menu or the right-click pop-up menu choose:
   Select > **Data Cells**
4. From the Format menu or the right-click pop-up menu choose **Cell Properties**.
5. Click the **Format Value** tab.
6. Type `0` in the Decimals field to hide all decimal points in this column.

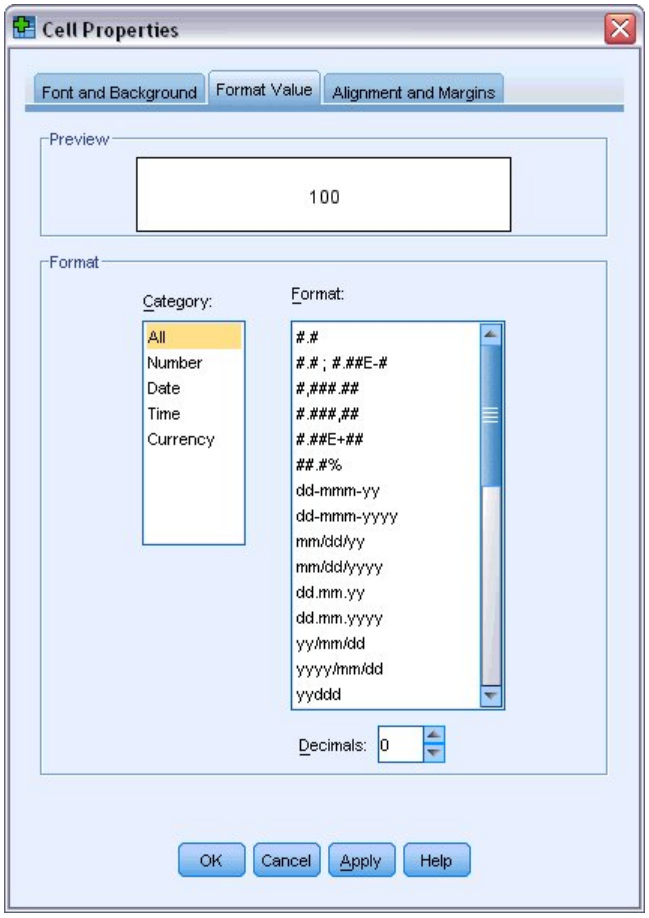


Figure 45. Cell Properties, Format Value tab

You can also change the data type and format in this dialog box.

7. Select the type that you want from the Category list, and then select the format for that type in the Format list.
8. Click **OK** or **Apply** to apply your changes.

Education Level

		Frequency	Percent	Cumulative Percent
Valid	Did not complete high school	1390	22	21.7
	High school degree	1936	30	52.0
	Some college	1360	21	73.2
	College degree	1355	21	94.4
	Post-undergraduate degree	359	6	100.0
	Total	6400	100	

Figure 46. Decimals hidden in Percent column

The decimals are now hidden in the *Percent* column.

TableLooks

The format of your tables is a critical part of providing clear, concise, and meaningful results. If your table is difficult to read, the information contained within that table may not be easily understood.

Using Predefined Formats

1. Double-click the *Marital status* table.
2. From the menus choose:

   Format > **TableLooks...**

   The TableLooks dialog box lists a variety of predefined styles. Select a style from the list to preview it in the Sample window on the right.

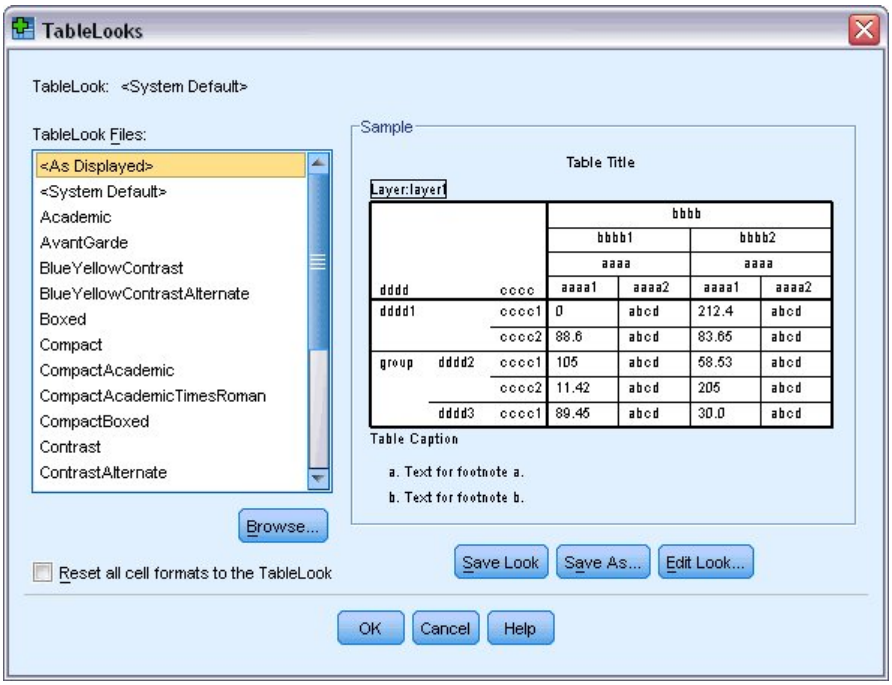


Figure 47. TableLooks dialog box

   You can use a style as is, or you can edit an existing style to better suit your needs.
3. To use an existing style, select one and click **OK**.

Customizing TableLook Styles

You can customize a format to fit your specific needs. Almost all aspects of a table can be customized, from the background color to the border styles.

1. Double-click the *Marital status* table.
2. From the menus choose:

   Format > **TableLooks...**
3. Select the style that is closest to the format you want and click **Edit Look**.
4. Click the **Cell Formats** tab to view the formatting options.

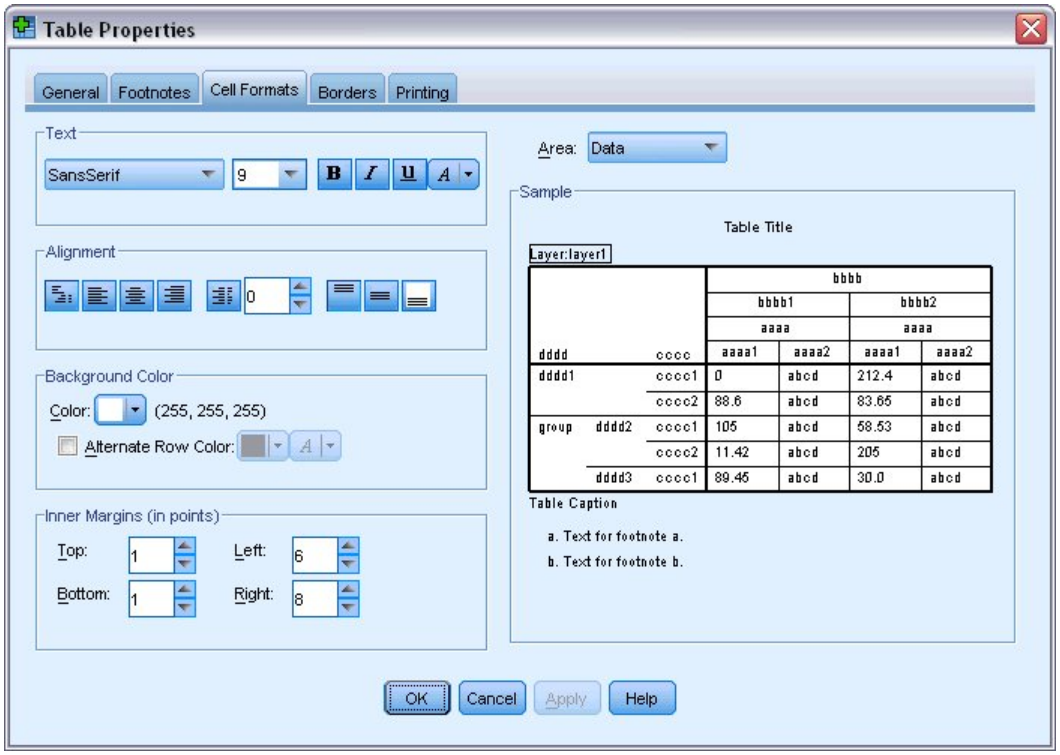


Figure 48. Table Properties dialog box

The formatting options include font name, font size, style, and color. Additional options include alignment, text and background colors, and margin sizes.

The Sample window on the right provides a preview of how the formatting changes affect your table. Each area of the table can have different formatting styles. For example, you probably wouldn't want the title to have the same style as the data. To select a table area to edit, you can either choose the area by name in the Area drop-down list, or you can click the area that you want to change in the Sample window.

5. Select **Data** from the Area drop-down list.
6. Select a new color from the Background drop-down palette.
7. Then select a new text color.

   The Sample window shows the new style.

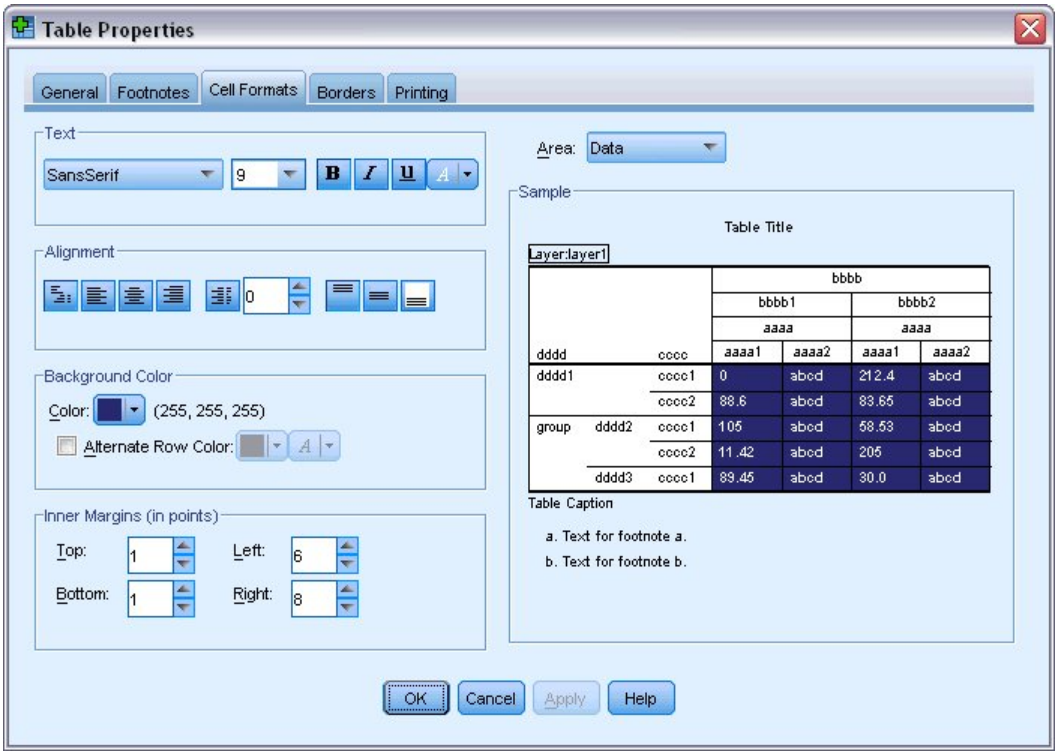


Figure 49. Changing table cell formats

8. Click **OK** to return to the TableLooks dialog box.
   You can save your new style, which allows you to apply it to future tables easily.
9. Click **Save As**.
10. Navigate to the target directory and enter a name for your new style in the File Name text box.
11. Click **Save**.
12. Click **OK** to apply your changes and return to the Viewer.

The table now contains the custom formatting that you specified.

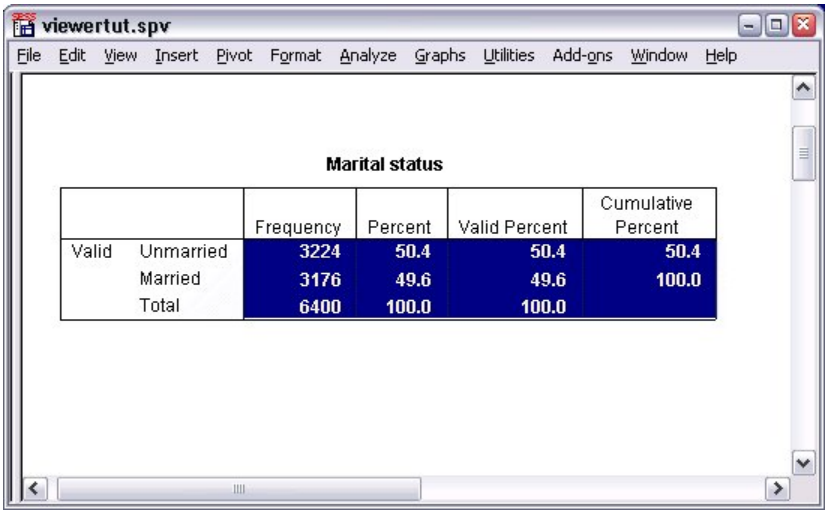


Figure 50. Custom TableLook

Changing the Default Table Formats

Although you can change the format of a table after it has been created, it may be more efficient to change the default TableLook so that you do not have to change the format every time you create a table.

To change the default TableLook style for your pivot tables, from the menus choose:

Edit > **Options...**

1. Click the **Pivot Tables** tab in the Options dialog box.

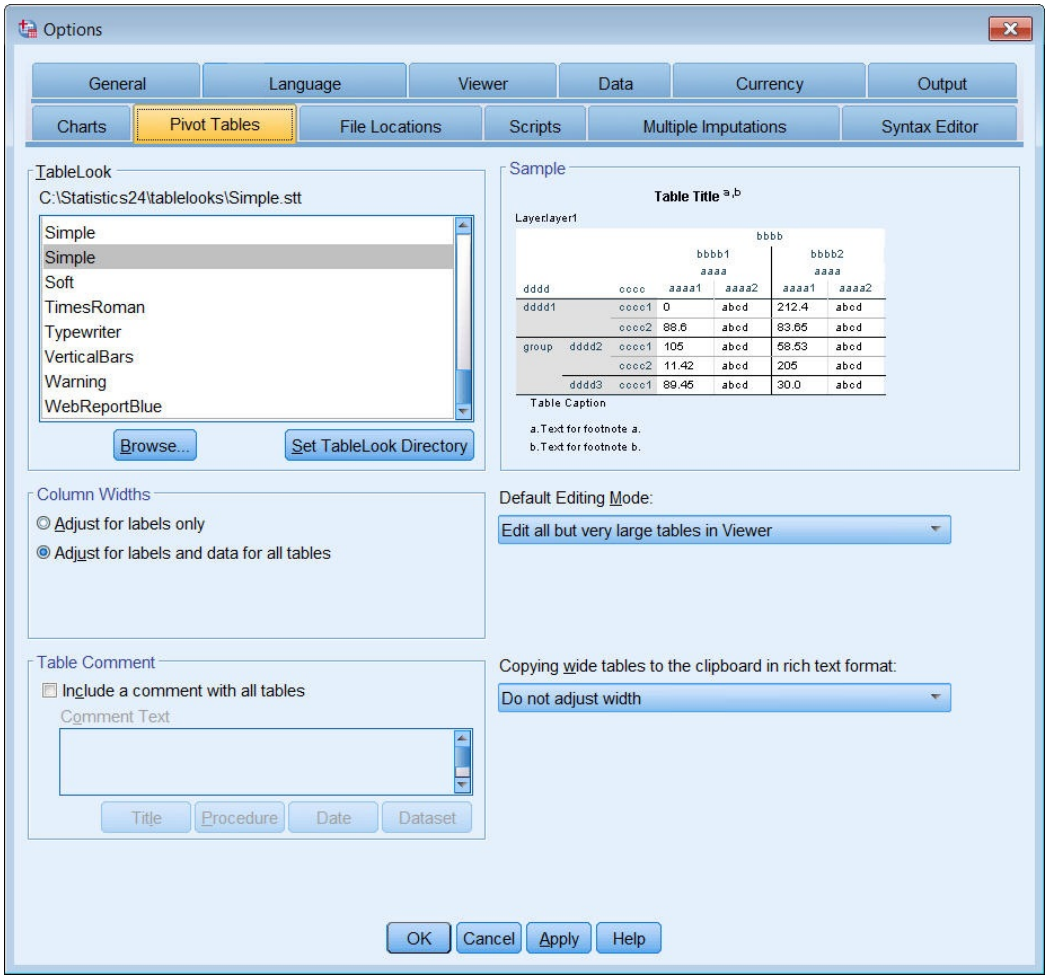


Figure 51. Options dialog box

2. Select the TableLook style that you want to use for all new tables.
   The Sample window on the right shows a preview of each TableLook.
3. Click **OK** to save your settings and close the dialog box.

All tables that you create after changing the default TableLook automatically conform to the new formatting rules.

Customizing the Initial Display Settings

The initial display settings include the alignment of objects in the Viewer, whether objects are shown or hidden by default, and the width of the Viewer window. To change these settings:

1. From the menus choose:

Edit > **Options...**

2. Click the **Viewer** tab.

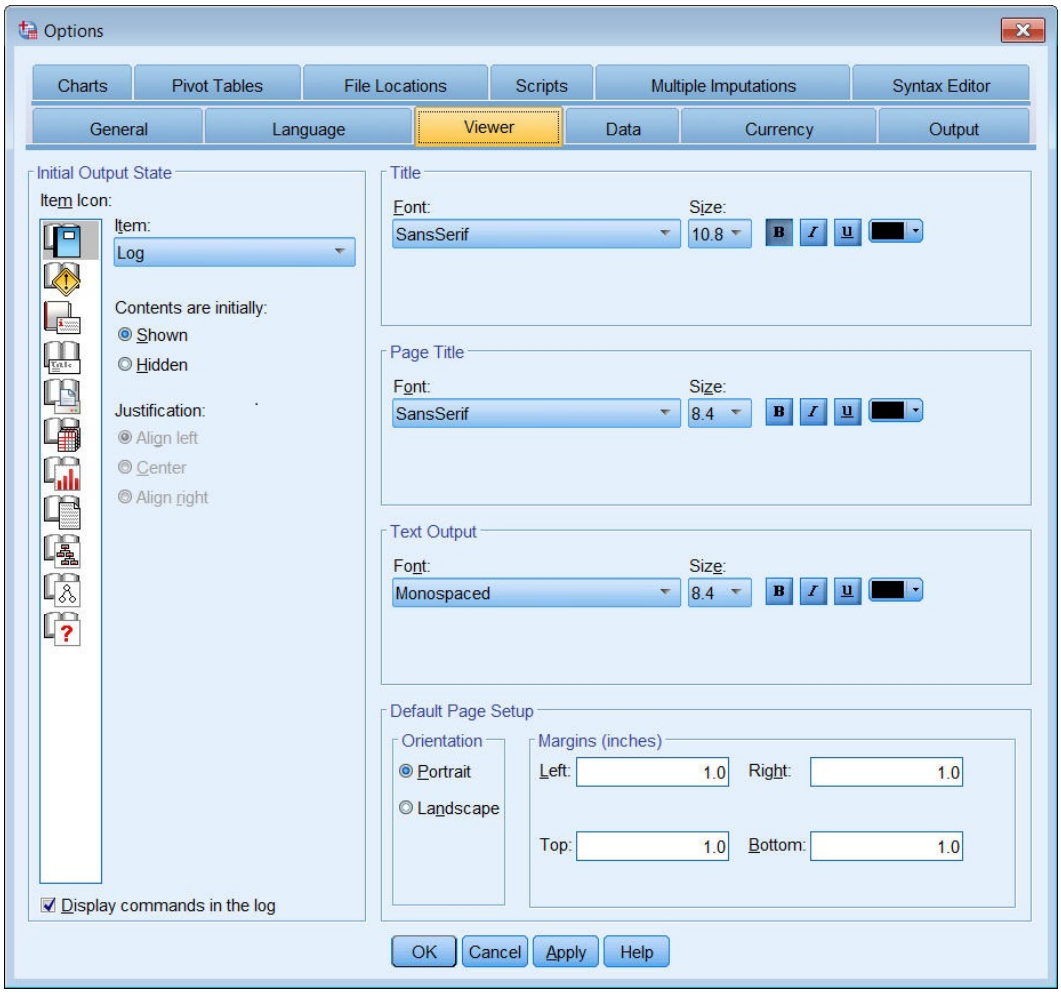


Figure 52. Viewer options

The settings are applied on an object-by-object basis. For example, you can customize the way charts are displayed without making any changes to the way tables are displayed. Simply select the object that you want to customize, and make the changes.

3. Click the **Title** icon to display its settings.
4. Click **Center** to display all titles in the (horizontal) center of the Viewer.

   You can also hide elements, such as the log and warning messages, that tend to clutter your output. Double-clicking on an icon automatically changes that object's display property.
5. Double-click the **Warnings** icon to hide warning messages in the output.
6. Click **OK** to save your changes and close the dialog box.

Displaying Variable and Value Labels

In most cases, displaying the labels for variables and values is more effective than displaying the variable name or the actual data value. There may be cases, however, when you want to display both the names and the labels.

1. From the menus choose:

   Edit > **Options...**

2. Click the **Output Labels** tab.

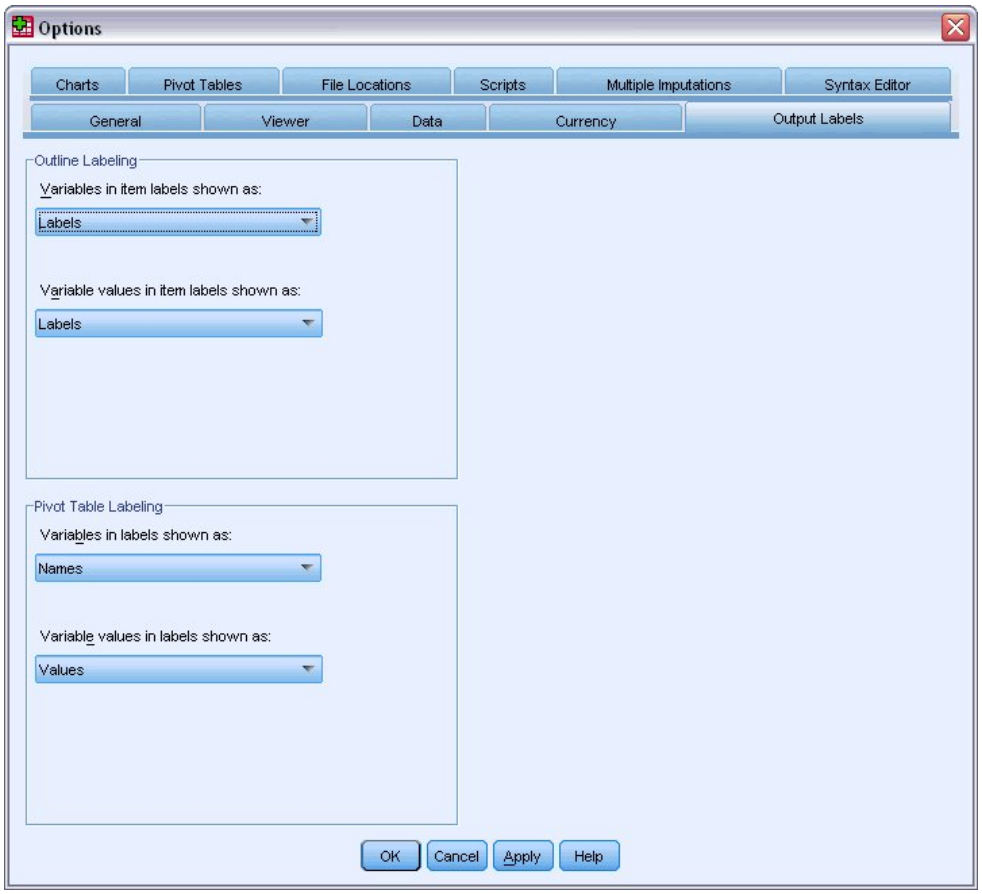


Figure 53. Pivot Table Labeling settings

You can specify different settings for the outline and contents panes. For example, to show labels in the outline and variable names and data values in the contents:

3. In the Pivot Table Labeling group, select **Names** from the Variables in Labels drop-down list to show variable names instead of labels.
4. Then, select **Values** from the Variable Values in Labels drop-down list to show data values instead of labels.

Subsequent tables produced in the session will reflect these changes.

marital

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3224	50.4	50.4	50.4
	1	3176	49.6	49.6	100.0
	Total	6400	100.0	100.0	

Figure 54. Variable names and values displayed

Using Results in Other Applications

Your results can be used in many applications. For example, you may want to include a table or chart in a presentation or report.

The following examples are specific to Microsoft Word, but they may work similarly in other word processing applications.

Pasting Results as Word Tables

You can paste pivot tables into Word as native Word tables. All table attributes, such as font sizes and colors, are retained. Because the table is pasted in the Word table format, you can edit it in Word just like any other table.

1. Click a table in the Viewer to select it.
2. From the menus choose:

   Edit > **Copy**
3. Open your word processing application.
4. From the word processor's menus choose:

   Edit > **Paste Special...**
5. Select **Formatted Text (RTF)** in the Paste Special dialog box.
6. Click **OK** to paste your results into the current document.

The table is now displayed in your document. You can apply custom formatting, edit the data, and resize the table to fit your needs.

Pasting Results as Text

Pivot tables can be copied to other applications as plain text. Formatting styles are not retained in this method, but you can edit the table data after you paste it into the target application.

1. Click a table in the Viewer to select it.
2. From the menus choose:

   Edit > **Copy**
3. Open your word processing application.
4. From the word processor's menus choose:

   Edit > **Paste Special...**
5. Select **Unformatted Text** in the Paste Special dialog box.
6. Click **OK** to paste your results into the current document.

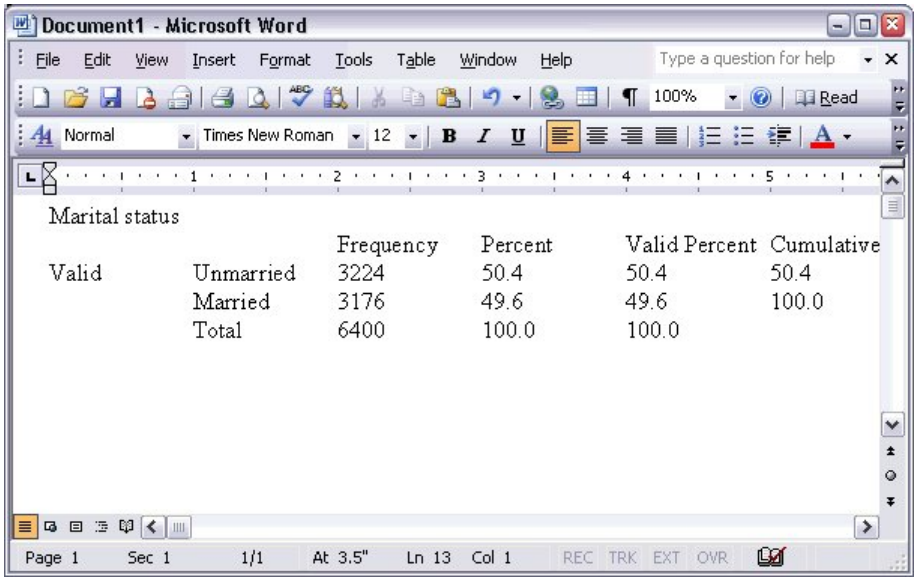


Figure 55. Pivot table displayed in Word

Each column of the table is separated by tabs. You can change the column widths by adjusting the tab stops in your word processing application.

Exporting Results to Microsoft Word, PowerPoint, and Excel Files

You can export results to a Microsoft Word , PowerPoint, or Excel file. You can export selected items or all items in the Viewer. This section uses the files *msouttut.spv* and *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.

Note: Export to PowerPoint is available only on Windows operating systems and is not available with the Student Version.

In the Viewer's outline pane, you can select specific items that you want to export or export all items or all visible items.

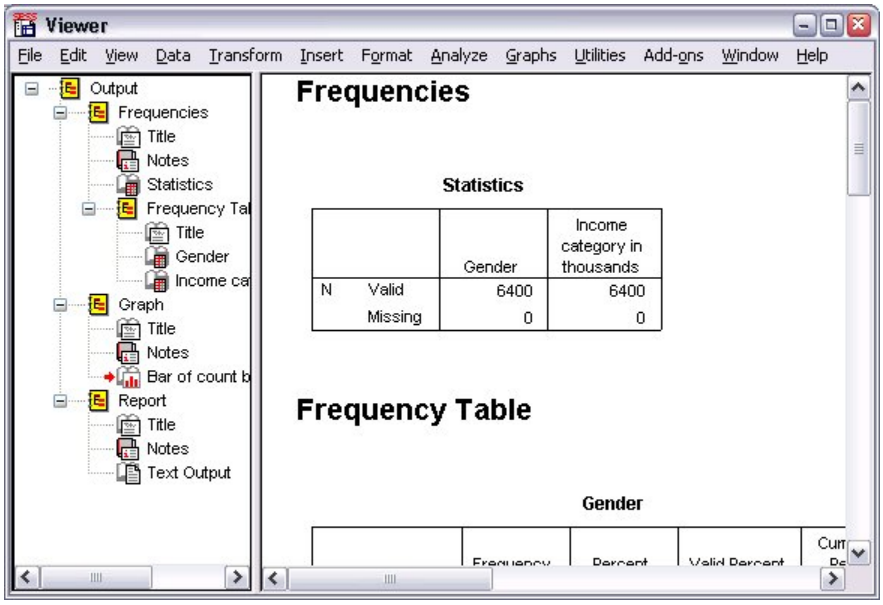


Figure 56. Viewer

1. From the Viewer menus choose:

   File > **Export...**

   Instead of exporting all objects in the Viewer, you can choose to export only visible objects (open books in the outline pane) or those that you selected in the outline pane. If you did not select any items in the outline pane, you do not have the option to export selected objects.

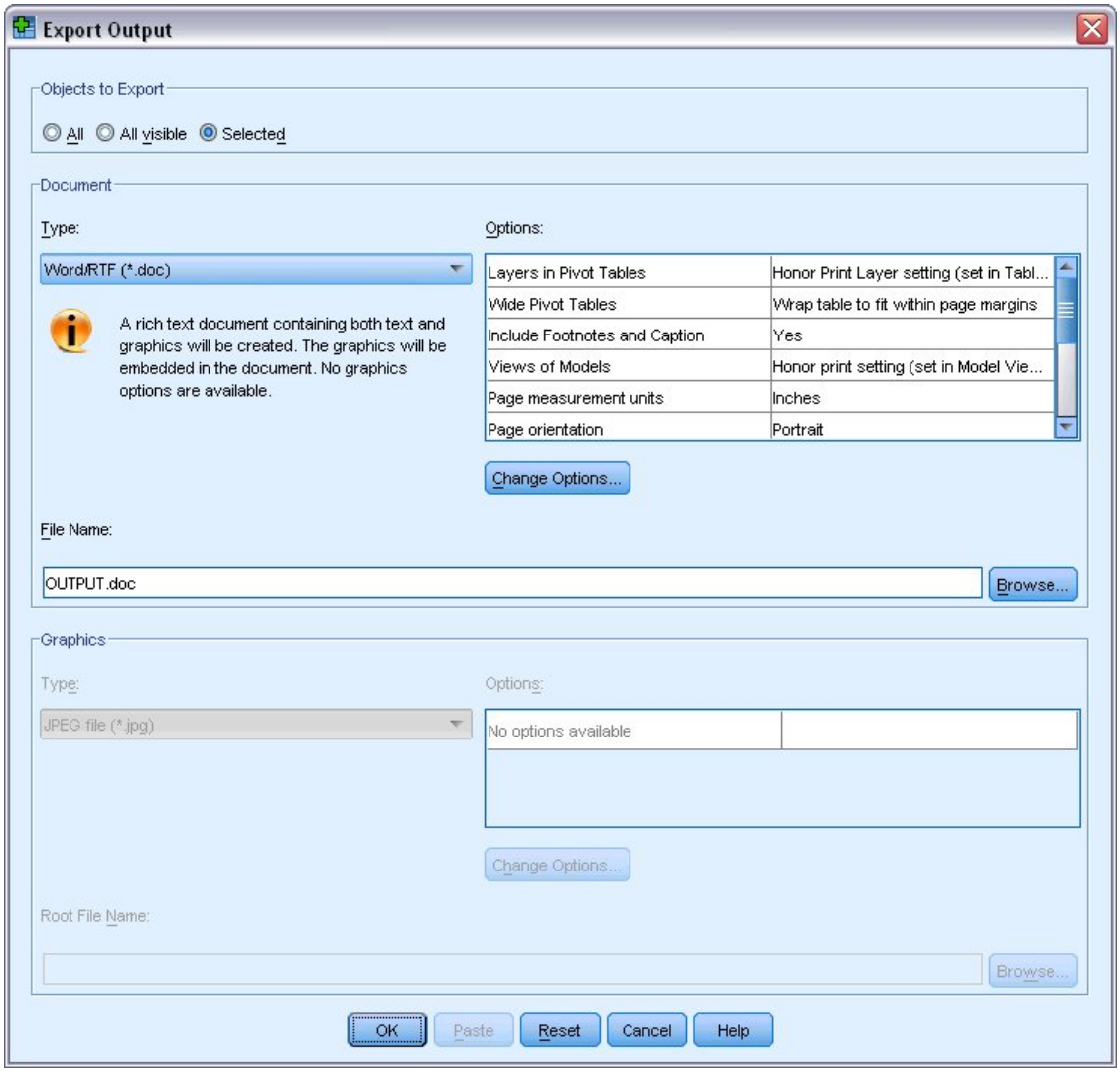


Figure 57. Export Output dialog box

2. In the Objects to Export group, select **All**.
3. From the Type drop-down list select **Word/RTF file (*.doc)**.
4. Click **OK** to generate the Word file.

When you open the resulting file in Word, you can see how the results are exported. Notes, which are not visible objects, appear in Word because you chose to export all objects.

Pivot tables become Word tables, with all of the formatting of the original pivot table retained, including fonts, colors, borders, and so on.

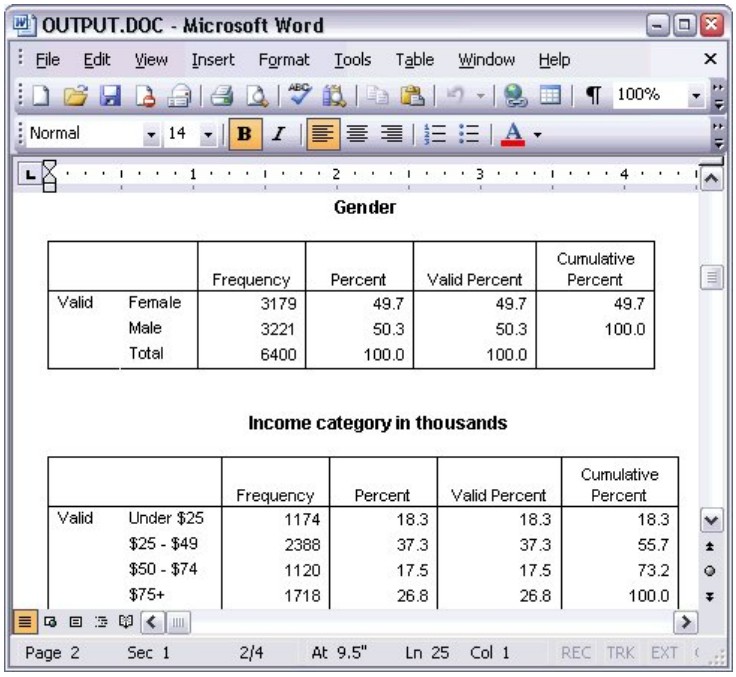


Figure 58. Pivot tables in Word

Charts are included in the Word document as graphic images.

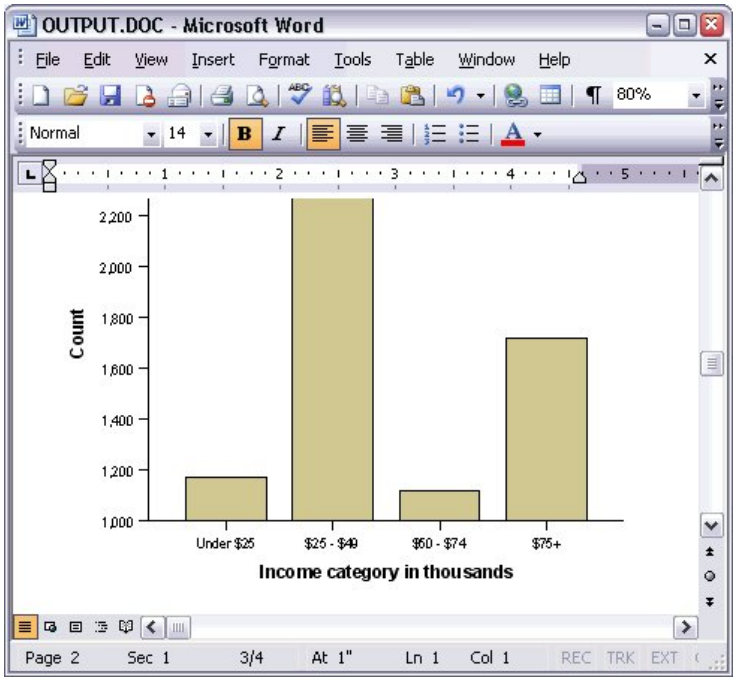


Figure 59. Charts in Word

Text output is displayed in the same font used for the text object in the Viewer. For proper alignment, text output should use a fixed-pitch (monospaced) font.

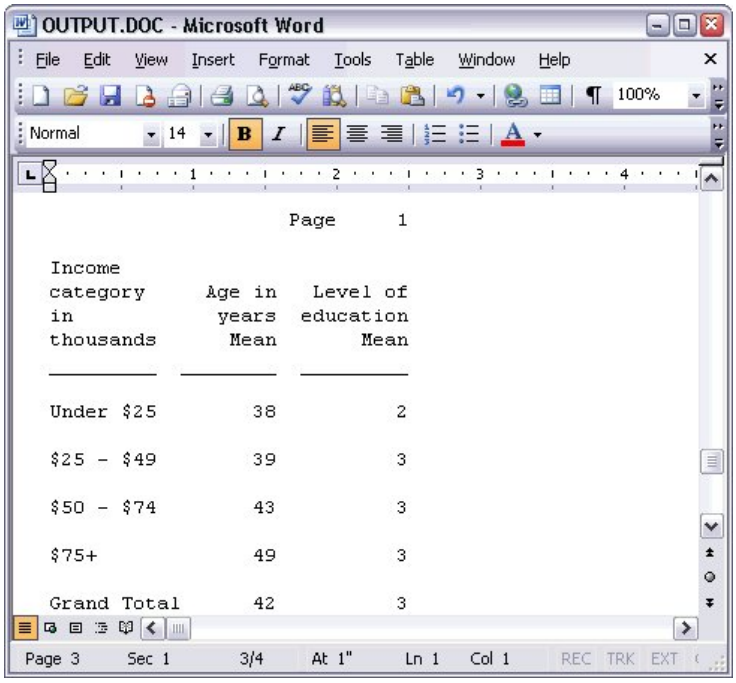


Figure 60. Text output in Word

If you export to a PowerPoint file, each exported item is placed on a separate slide. Pivot tables exported to PowerPoint become Word tables, with all of the formatting of the original pivot table, including fonts, colors, borders, and so on.

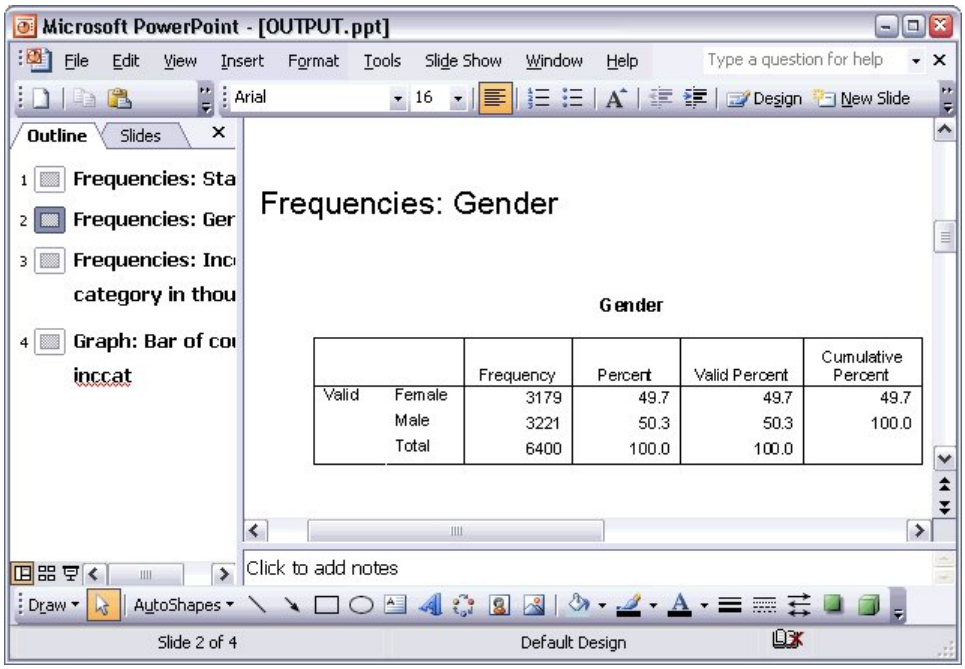


Figure 61. Pivot tables in PowerPoint

Charts selected for export to PowerPoint are embedded in the PowerPoint file.

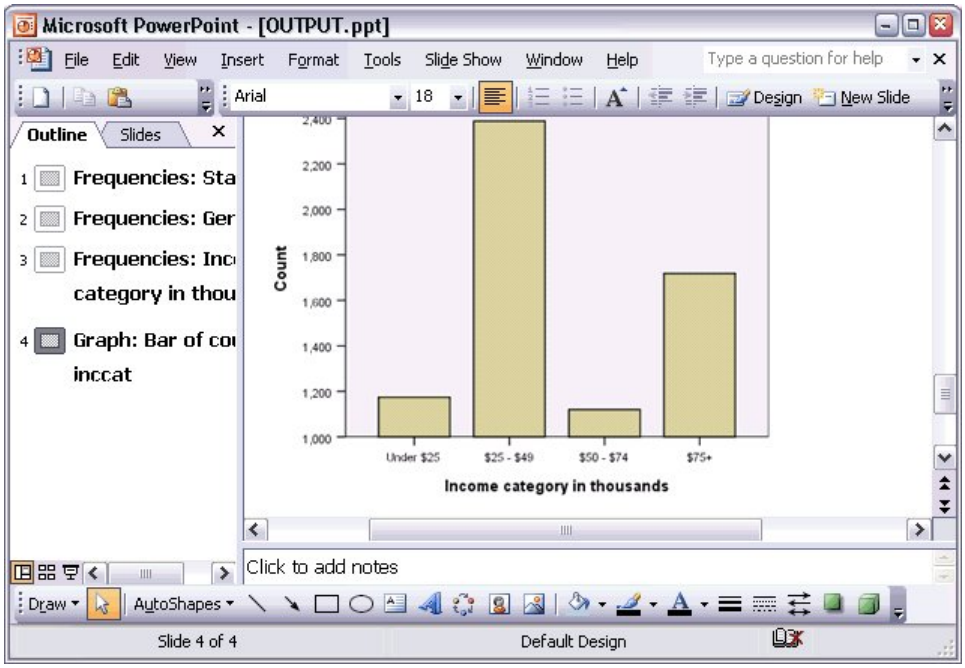
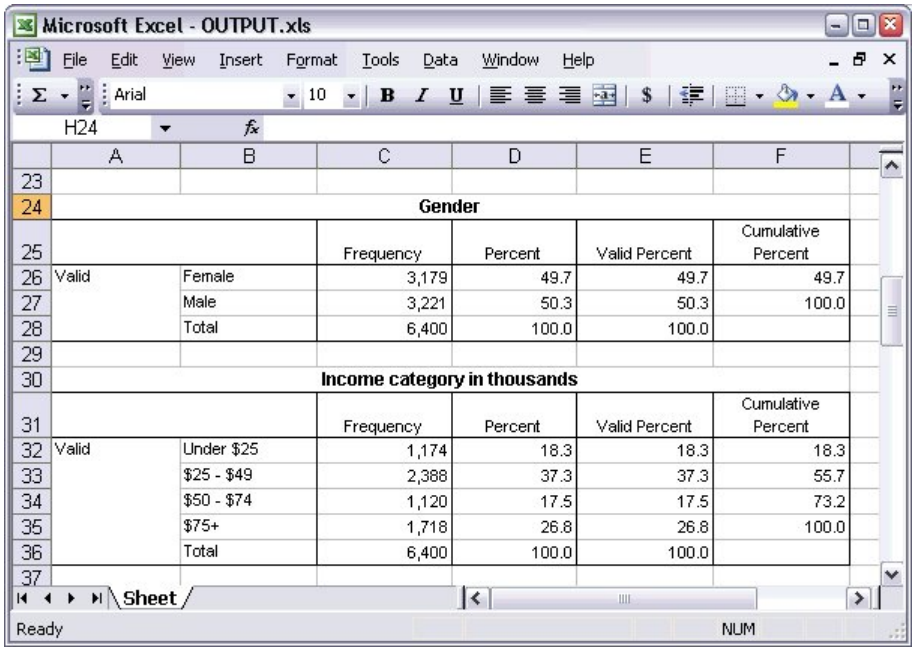


Figure 62. Charts in PowerPoint

Note: Export to PowerPoint is available only on Windows operating systems and is not available with the Student Version.

If you export to an Excel file, results are exported differently.



	A	B	C	D	E	F
23						
24	**Gender**					
25			Frequency	Percent	Valid Percent	Cumulative Percent
26	Valid	Female	3,179	49.7	49.7	49.7
27		Male	3,221	50.3	50.3	100.0
28		Total	6,400	100.0	100.0	
29						
30	**Income category in thousands**					
31			Frequency	Percent	Valid Percent	Cumulative Percent
32	Valid	Under $25	1,174	18.3	18.3	18.3
33		$25 - $49	2,388	37.3	37.3	55.7
34		$50 - $74	1,120	17.5	17.5	73.2
35		$75+	1,718	26.8	26.8	100.0
36		Total	6,400	100.0	100.0	
37						

Figure 63. Output.xls in Excel

Pivot table rows, columns, and cells become Excel rows, columns, and cells.

Each line in the text output is a row in the Excel file, with the entire contents of the line contained in a single cell.

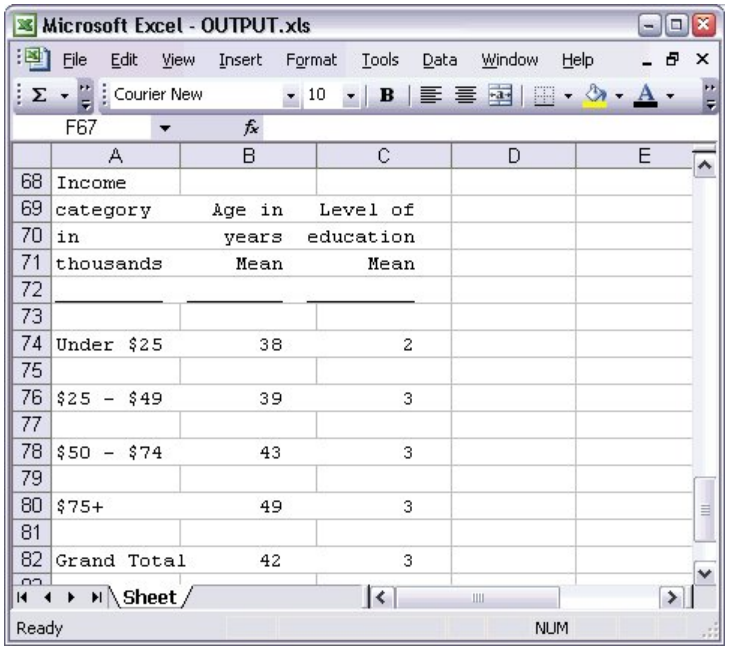


Figure 64. Text output in Excel

Exporting Results to PDF

You can export all or selected items in the Viewer to a PDF (portable document format) file.

1. From the menus in the Viewer window that contains the result you want to export to PDF choose:
   File > **Export...**
2. In the Export Output dialog box, from the Export Format File Type drop-down list choose **Portable Document Format**.

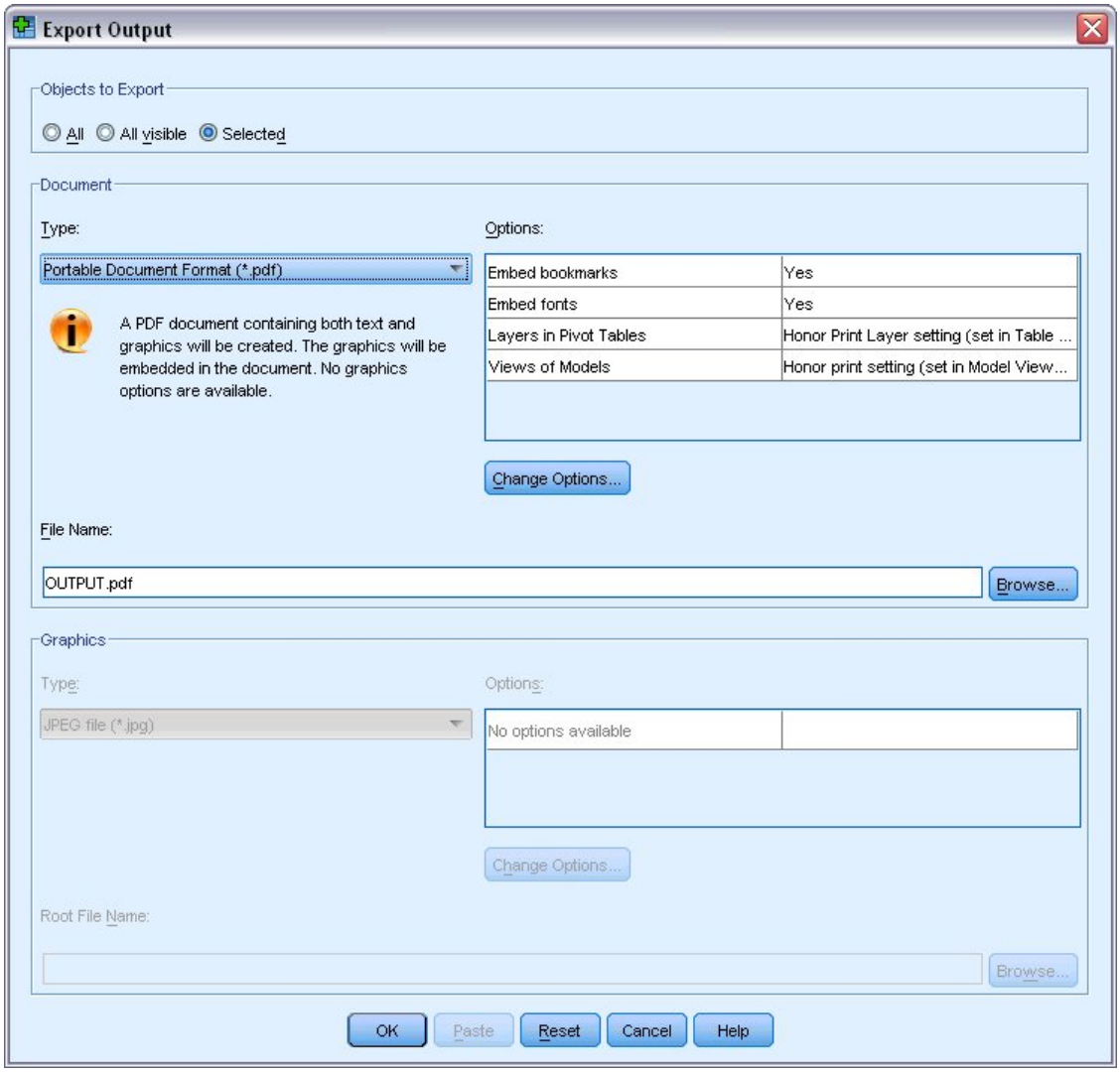


Figure 65. Export Output dialog box

- The outline pane of the Viewer document is converted to bookmarks in the PDF file for easy navigation.
- Page size, orientation, margins, content and display of page headers and footers, and printed chart size in PDF documents are controlled by page setup options (File menu, Page Setup in the Viewer window).
- The resolution (DPI) of the PDF document is the current resolution setting for the default or currently selected printer (which can be changed using Page Setup). The maximum resolution is 1200 DPI. If the printer setting is higher, the PDF document resolution will be 1200 DPI. *Note*: High-resolution documents may yield poor results when printed on lower-resolution printers.

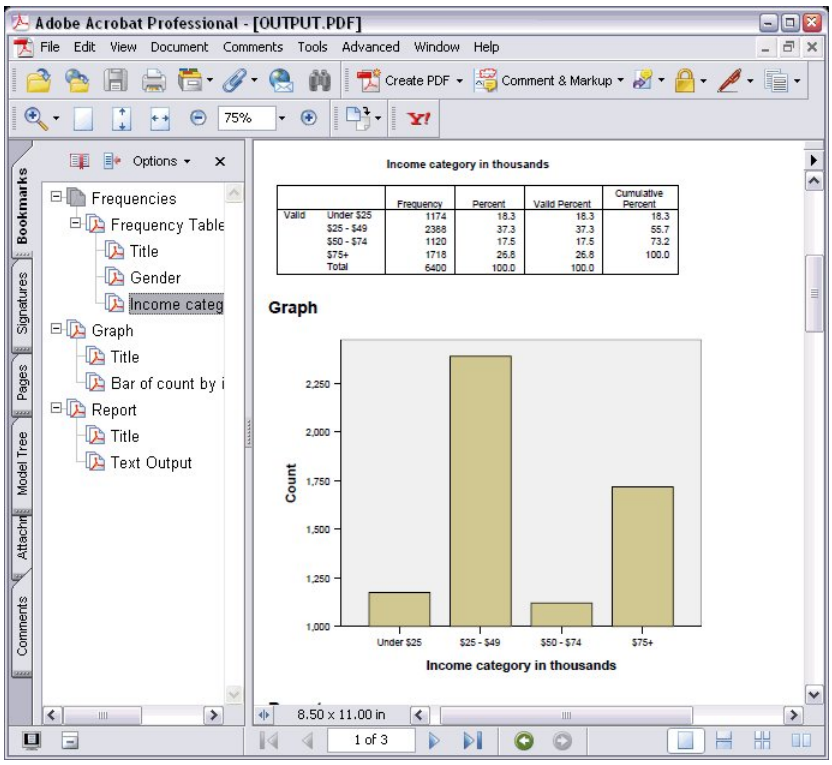


Figure 66. PDF file with bookmarks

Exporting Results to HTML

You can also export results to HTML (hypertext markup language). When saving as HTML, all non-graphic output is exported into a single HTML file.

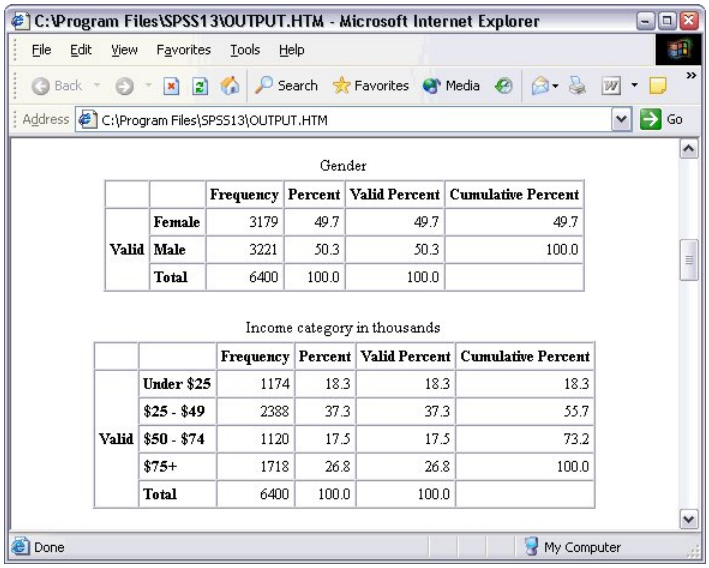


Figure 67. Output.htm in Web browser

When you export to HTML, charts can be exported as well, but not to a single file.

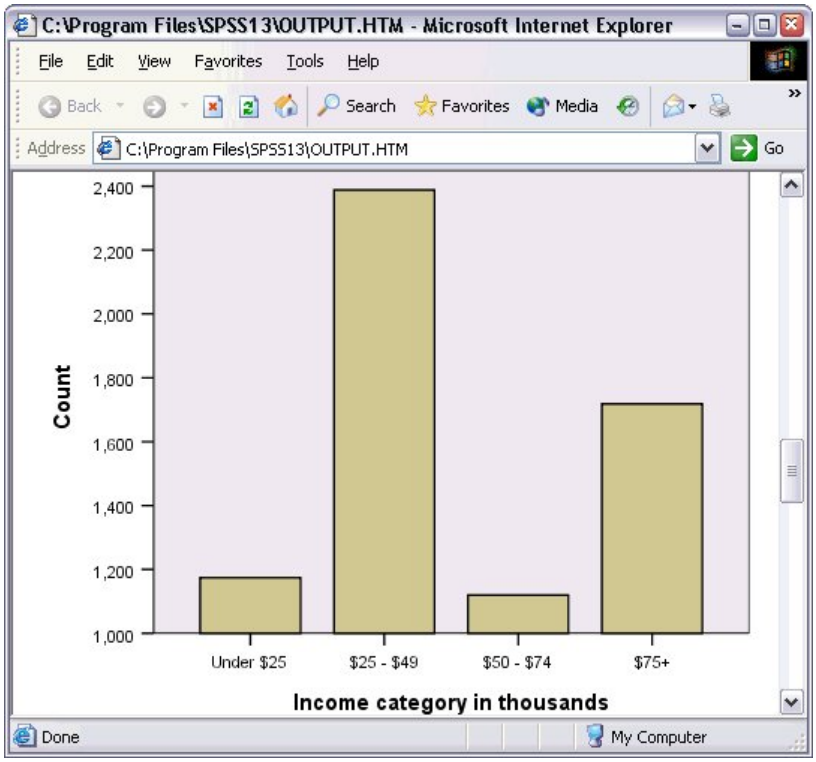


Figure 68. Chart in HTML

Each chart will be saved as a file in a format that you specify, and references to these graphics files will be placed in the HTML. There is also an option to export all charts (or selected charts) to separate graphics files.

Chapter 7. Working with Syntax

You can save and automate many common tasks by using the powerful command language. It also provides some functionality not found in the menus and dialog boxes. Most commands are accessible from the menus and dialog boxes. However, some commands and options are available only by using the command language. The command language also allows you to save your jobs in a syntax file so that you can repeat your analysis at a later date.

A command syntax file is simply a text file that contains IBM SPSS Statistics syntax commands. You can open a syntax window and type commands directly, but it is often easier to let the dialog boxes do some or all of the work for you.

The examples in this chapter use the data file *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.

Note: Command syntax is not available with the Student Version.

Pasting Syntax

The easiest way to create syntax is to use the Paste button located on most dialog boxes.

1. Open the data file *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.
2. From the menus choose:

   Analyze > **Descriptive Statistics** > **Frequencies...**
3. Select *Marital status [marital]* and move it into the Variable(s) list.
4. Click **Charts**.
5. In the Charts dialog box, select **Bar charts**.
6. In the Chart Values group, select **Percentages**.
7. Click **Continue**.Click **Paste** to copy the syntax created as a result of the dialog box selections to the Syntax Editor.

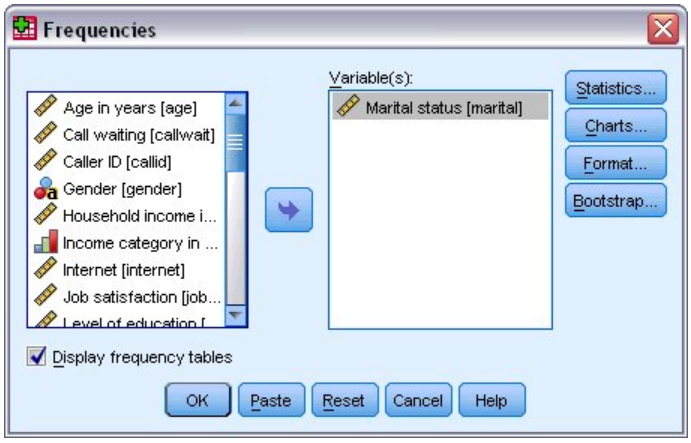


Figure 69. Frequencies dialog box

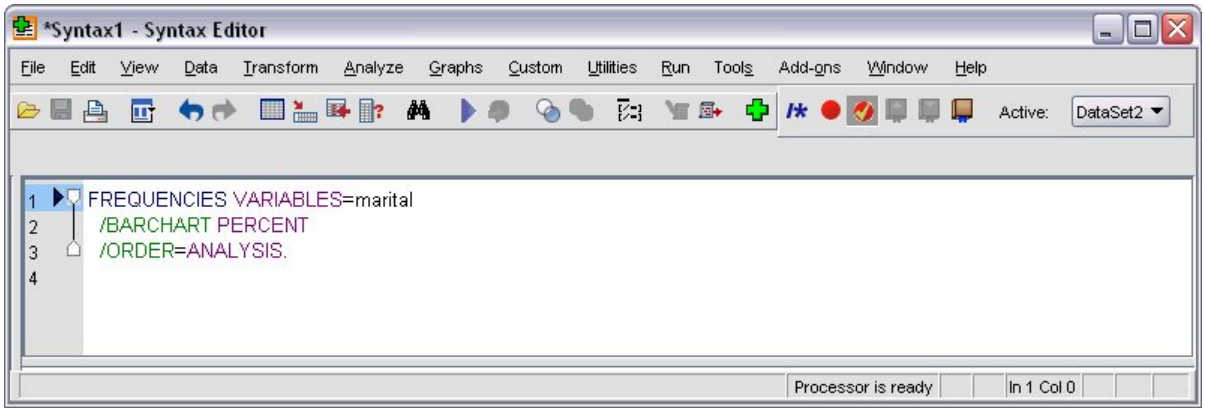


Figure 70. Frequencies syntax

8. To run the syntax currently displayed, from the menus choose:
   Run > **Selection**

Editing Syntax

In the syntax window, you can edit the syntax. For example, you could change the subcommand `/BARCHART` to display frequencies instead of percentages. (A subcommand is indicated by a slash.) If you know the keyword for displaying frequencies you can enter it directly. If you don't know the keyword, you can obtain a list of the available keywords for the subcommand by positioning the cursor anywhere following the subcommand name and pressing Ctrl+Spacebar. This displays the auto-completion control for the subcommand.

Delete the keyword `PERCENT` from the `BARCHART` subcommand.

Press Ctrl-Spacebar.

1. click the item labelled **FREQ** for frequencies. Clicking on an item in the auto-completion control will insert it at the current cursor position.

   By default, the auto-completion control will prompt you with a list of available terms as you type. For example, you'd like to include a pie chart along with the bar chart. The pie chart is specified with a separate subcommand.
2. Press Enter after the **FREQ** keyword and type a forward slash to indicate the start of a subcommand.

The Syntax Editor prompts you with the list of subcommands for the current command.

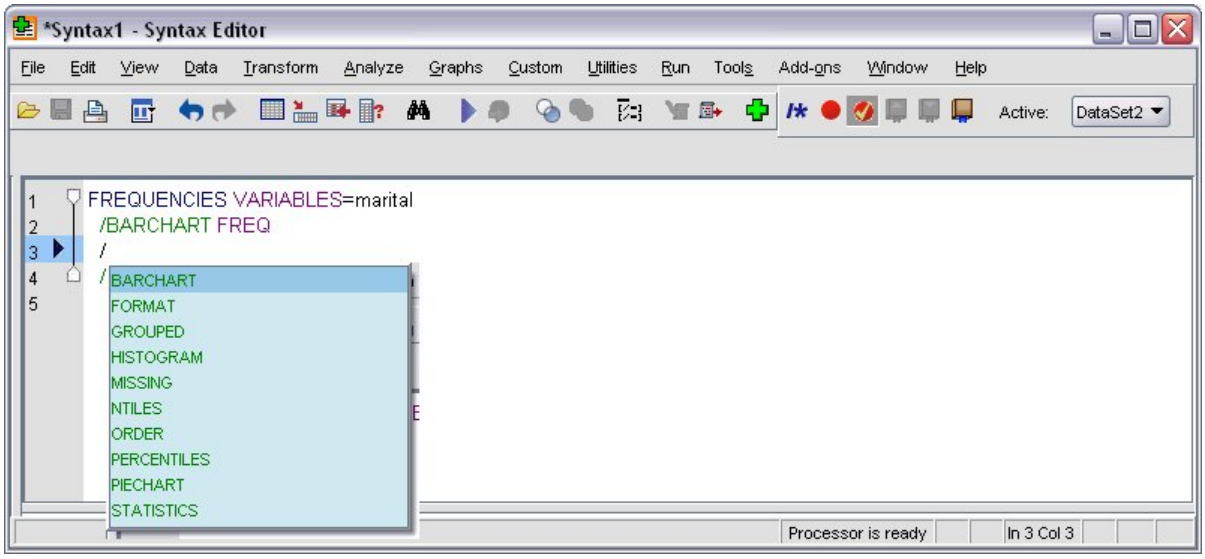


Figure 71. Auto-completion control displaying subcommands

To obtain more detailed help for the current command, press the F1 key. This takes you directly to the command syntax reference information for the current command.

You may have noticed that text displayed in the syntax window is colored. Color coding allows you to quickly identify unrecognized terms, since only recognized terms are colored. For example, you misspell the `FORMAT` subcommand as `FRMAT`. Subcommands are colored green by default, but the text `FRMAT` will appear uncolored since it is not recognized.

Opening and Running a Syntax File

1. To open a saved syntax file, from the menus choose:
   File > **Open** > **Syntax...**
   A standard dialog box for opening files is displayed.
2. Select a syntax file. If no syntax files are displayed, make sure **Syntax (*.sps)** is selected as the file type you want to view.
3. Click **Open**.
4. Use the Run menu in the syntax window to run the commands.

If the commands apply to a specific data file, the data file must be opened before running the commands, or you must include a command that opens the data file. You can paste this type of command from the dialog boxes that open data files.

Using Breakpoints

Breakpoints allow you to stop execution of command syntax at specified points within the syntax and continue execution when ready. This allows you to view output or data at an intermediate point in a syntax job, or to run command syntax that displays information about the current state of the data, such as `FREQUENCIES`. Breakpoints can only be set at the level of a command, not on specific lines within a command.

To insert a breakpoint on a command:

1. Click anywhere in the region to the left of the text associated with the command.

The breakpoint is represented as a red circle in the region to the left of the command text and on the same line as the command name regardless of where you clicked.

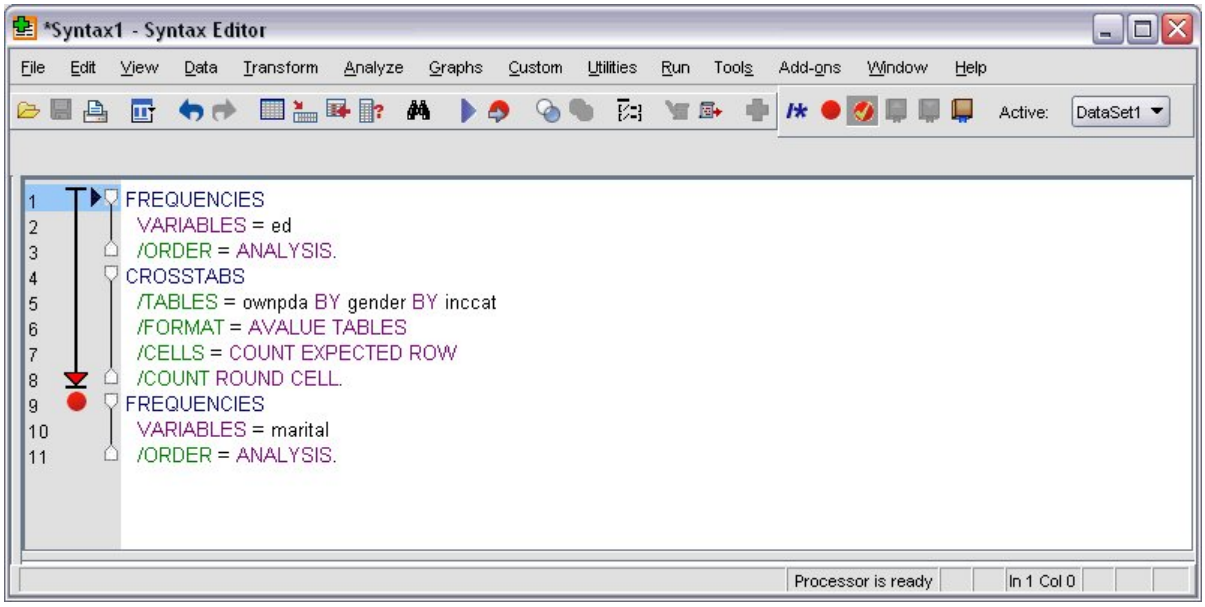


Figure 72. Execution stopped at a breakpoint

When you run command syntax containing breakpoints, execution stops prior to each command containing a breakpoint.

The downward pointing arrow to the left of the command text shows the progress of the syntax run. It spans the region from the first command run through the last command run and is particularly useful when running command syntax containing breakpoints.

To resume execution following a breakpoint:

2. From the menus in the Syntax Editor choose:

   Run > **Continue**

Chapter 8. Modifying Data Values

The data you start with may not always be organized in the most useful manner for your analysis or reporting needs. For example, you may want to:

- Create a categorical variable from a scale variable.
- Combine several response categories into a single category.
- Create a new variable that is the computed difference between two existing variables.
- Calculate the length of time between two dates.

This chapter uses the data file *demo.sav*. See the topic Chapter 10, “Sample Files,” on page 83 for more information.

Creating a Categorical Variable from a Scale Variable

Several categorical variables in the data file *demo.sav* are, in fact, derived from scale variables in that data file. For example, the variable *inccat* is simply *income* grouped into four categories. This categorical variable uses the integer values 1–4 to represent the following income categories (in thousands): less than $25, $25–$49, $50–$74, and $75 or higher.

To create the categorical variable *inccat*:

1. From the menus in the Data Editor window choose:

   Transform > **Visual Binning...**

   In the initial Visual Binning dialog box, you select the scale and/or ordinal variables for which you want to create new, binned variables. **Binning** means taking two or more contiguous values and grouping them into the same category.

   Since Visual Binning relies on actual values in the data file to help you make good binning choices, it needs to read the data file first. Since this can take some time if your data file contains a large number of cases, this initial dialog box also allows you to limit the number of cases to read ("scan"). This is not necessary for our sample data file. Even though it contains more than 6,000 cases, it does not take long to scan that number of cases.
2. Drag and drop *Household income in thousands [income]* from the Variables list into the Variables to Bin list, and then click **Continue**.

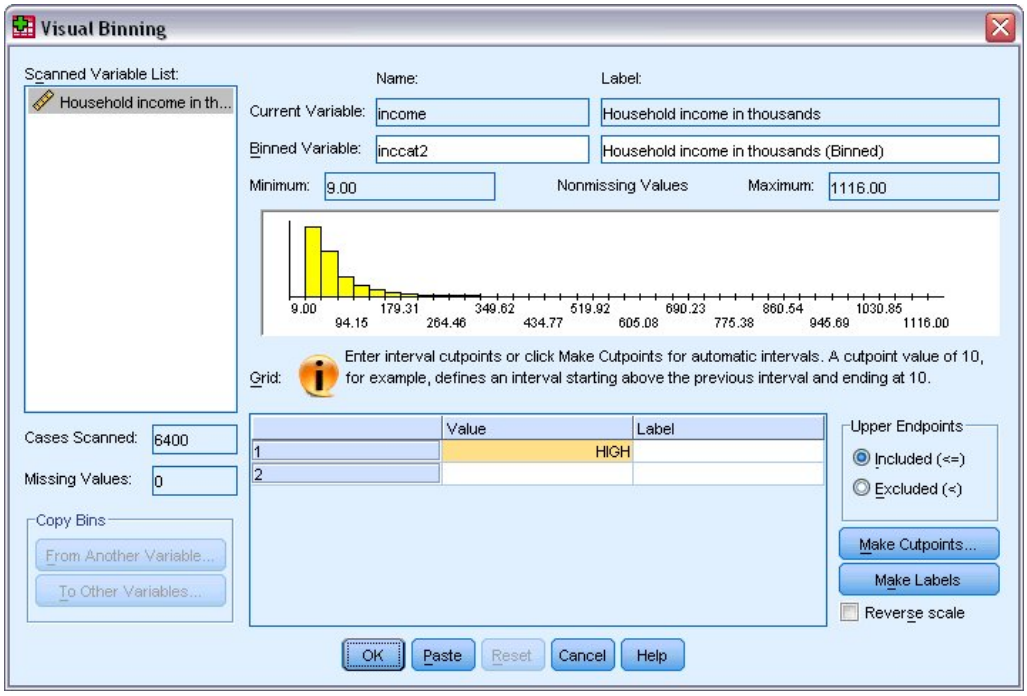


Figure 73. Main Visual Binning dialog box

3. In the main Visual Binning dialog box, select *Household income in thousands [income]* in the Scanned Variable List.

   A histogram displays the distribution of the selected variable (which in this case is highly skewed).
4. Enter `inccat2` for the new binned variable name and `Income category [in thousands]` for the variable label.
5. Click **Make Cutpoints**.
6. Select **Equal Width Intervals.**
7. Enter `25` for the first cutpoint location, `3` for the number of cutpoints, and `25` for the width.

   The number of binned categories is one greater than the number of cutpoints. So in this example, the new binned variable will have four categories, with the first three categories each containing ranges of 25 (thousand) and the last one containing all values above the highest cutpoint value of 75 (thousand).
8. Click **Apply**.

   The values now displayed in the grid represent the defined cutpoints, which are the upper endpoints of each category. Vertical lines in the histogram also indicate the locations of the cutpoints.

   By default, these cutpoint values are included in the corresponding categories. For example, the first value of 25 would include all values less than or equal to 25. But in this example, we want categories that correspond to less than 25, 25–49, 50–74, and 75 or higher.
9. In the Upper Endpoints group, select **Excluded (<).**
10. Then click **Make Labels**.

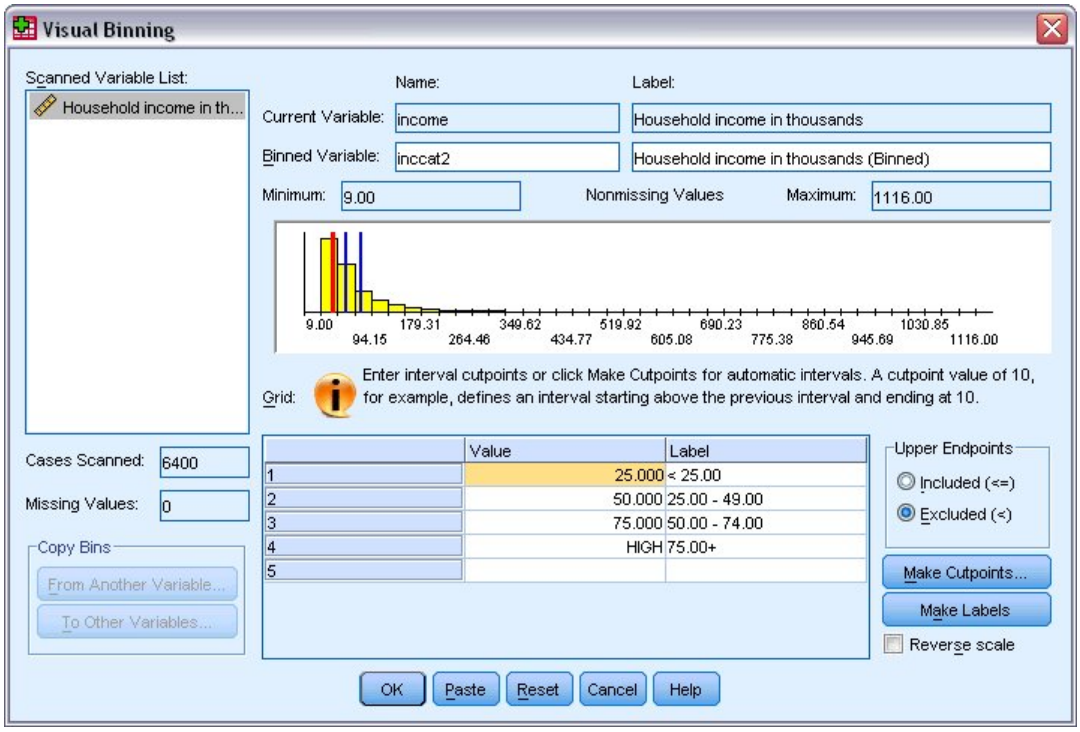


Figure 74. Automatically generated value labels

This automatically generates descriptive value labels for each category. Since the actual values assigned to the new binned variable are simply sequential integers starting with 1, the value labels can be very useful.

You can also manually enter or change cutpoints and labels in the grid, change cutpoint locations by dragging and dropping the cutpoint lines in the histogram, and delete cutpoints by dragging cutpoint lines off of the histogram.

11. Click **OK** to create the new, binned variable.

The new variable is displayed in the Data Editor. Since the variable is added to the end of the file, it is displayed in the far right column in Data View and in the last row in Variable View.

Computing New Variables

Using a wide variety of mathematical functions, you can compute new variables based on highly complex equations. In this example, however, we will simply compute a new variable that is the difference between the values of two existing variables.

The data file *demo.sav* contains a variable for the respondent's current age and a variable for the number of years at current job. It does not, however, contain a variable for the respondent's age at the time he or she started that job. We can create a new variable that is the computed difference between current age and number of years at current job, which should be the approximate age at which the respondent started that job.

1. From the menus in the Data Editor window choose:

   Transform > **Compute Variable...**

2. For Target Variable, enter `jobstart`.
3. Select *Age in years [age]* in the source variable list and click the arrow button to copy it to the Numeric Expression text box.

4. Click the minus (–) button on the calculator pad in the dialog box (or press the minus key on the keyboard).
5. Select *Years with current employer [employ]* and click the arrow button to copy it to the expression.

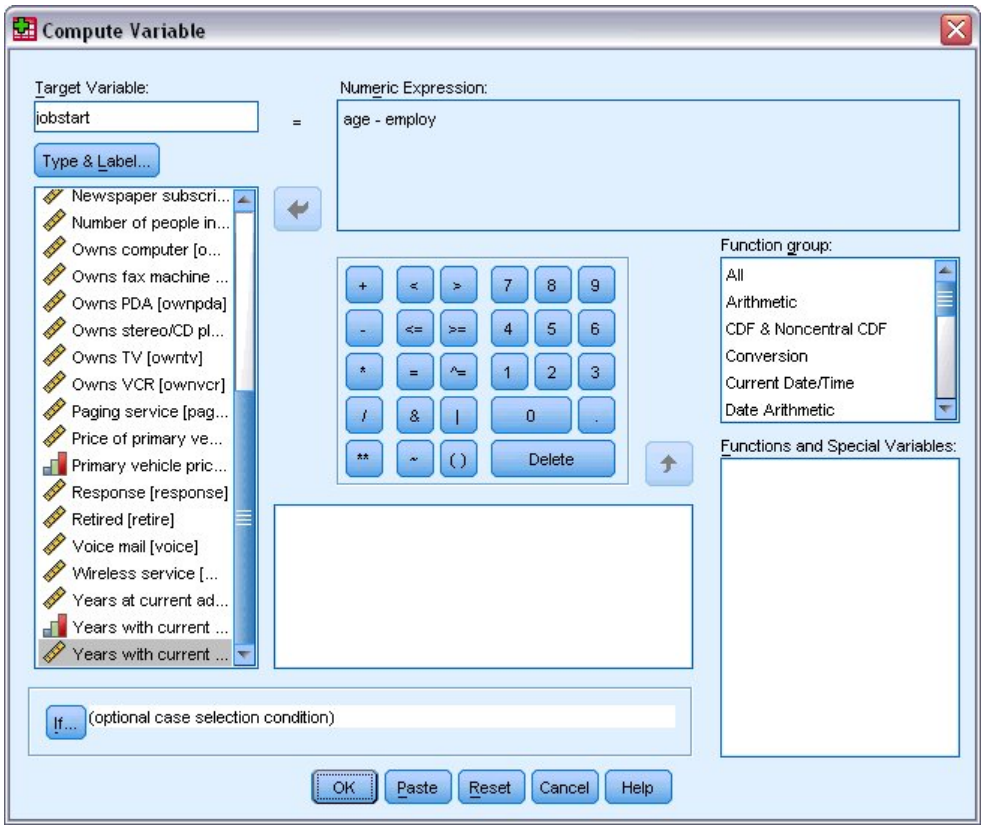


Figure 75. Compute Variable dialog box

Note: Be careful to select the correct employment variable. There is also a recoded categorical version of the variable, which is *not* what you want. The numeric expression should be *age–employ*, not *age–empcat*.

6. Click **OK** to compute the new variable.

The new variable is displayed in the Data Editor. Since the variable is added to the end of the file, it is displayed in the far right column in Data View and in the last row in Variable View.

Using Functions in Expressions

You can also use predefined functions in expressions. More than 70 built-in functions are available, including:

- Arithmetic functions
- Statistical functions
- Distribution functions
- Logical functions
- Date and time aggregation and extraction functions
- Missing-value functions
- Cross-case functions
- String functions

Functions are organized into logically distinct groups, such as a group for arithmetic operations and another for computing statistical metrics. For convenience, a number of commonly used system variables, such as *$TIME* (current date and time), are also included in appropriate function groups.

Pasting a Function into an Expression

To paste a function into an expression:

1. Position the cursor in the expression at the point where you want the function to appear.
2. Select the appropriate group from the Function group list. The group labeled **All** provides a listing of all available functions and system variables.
3. Double-click the function in the Functions and Special Variables list (or select the function and click the arrow adjacent to the Function group list).

The function is inserted into the expression. If you highlight part of the expression and then insert the function, the highlighted portion of the expression is used as the first argument in the function.

Editing a Function in an Expression

The function is not complete until you enter the arguments, represented by question marks in the pasted function. The number of question marks indicates the minimum number of arguments required to complete the function.

1. Highlight the question mark(s) in the pasted function.
2. Enter the arguments. If the arguments are variable names, you can paste them from the variable list.

Using Conditional Expressions

You can use conditional expressions (also called logical expressions) to apply transformations to selected subsets of cases. A conditional expression returns a value of true, false, or missing for each case. If the result of a conditional expression is true, the transformation is applied to that case. If the result is false or missing, the transformation is not applied to the case.

To specify a conditional expression:

1. Click **If** in the Compute Variable dialog box. This opens the If Cases dialog box.
2. Select **Include if case satisfies condition**.
3. Enter the conditional expression.

Most conditional expressions contain at least one relational operator, as in:

```
age>=21
```

or

```
income*3<100
```

In the first example, only cases with a value of 21 or greater for *Age [age]* are selected. In the second example, *Household income in thousands [income]* multiplied by 3 must be less than 100 for a case to be selected.

You can also link two or more conditional expressions using logical operators, as in:

```
age>=21 | ed>=4
```

or

```
income*3<100 & ed=5
```

In the first example, cases that meet either the *Age [age]* condition or the *Level of education [ed]* condition are selected. In the second example, both the *Household income in thousands [income]* and *Level of education [ed]* conditions must be met for a case to be selected.

Working with Dates and Times

A number of tasks commonly performed with dates and times can be easily accomplished using the Date and Time Wizard. Using this wizard, you can:

- Create a date/time variable from a string variable containing a date or time.
- Construct a date/time variable by merging variables containing different parts of the date or time.
- Add or subtract values from date/time variables, including adding or subtracting two date/time variables.
- Extract a part of a date or time variable; for example, the day of month from a date/time variable which has the form mm/dd/yyyy.

The examples in this section use the data file *upgrade.sav*. See the topic Chapter 10, “Sample Files,” on page 83 for more information.

To use the Date and Time Wizard:

1. From the menus choose:

Transform > **Date and Time Wizard...**

Figure 76. Date and Time Wizard introduction screen

The introduction screen of the Date and Time Wizard presents you with a set of general tasks. Tasks that do not apply to the current data are disabled. For example, the data file *upgrade.sav* doesn't contain any string variables, so the task to create a date variable from a string is disabled.

If you're new to dates and times in IBM SPSS Statistics, you can select **Learn how dates and times are represented** and click **Next**. This leads to a screen that provides a brief overview of date/time variables and a link, through the Help button, to more detailed information.

Calculating the Length of Time between Two Dates

One of the most common tasks involving dates is calculating the length of time between two dates. As an example, consider a software company interested in analyzing purchases of upgrade licenses by determining the number of years since each customer last purchased an upgrade. The data file *upgrade.sav* contains a variable for the date on which each customer last purchased an upgrade but not the number of years since that purchase. A new variable that is the length of time in years between the date of the last upgrade and the date of the next product release will provide a measure of this quantity.

To calculate the length of time between two dates:

1. Select **Calculate with dates and times** on the introduction screen of the Date and Time Wizard and click **Next**.
2. Select **Calculate the number of time units between two dates** and click **Next**.

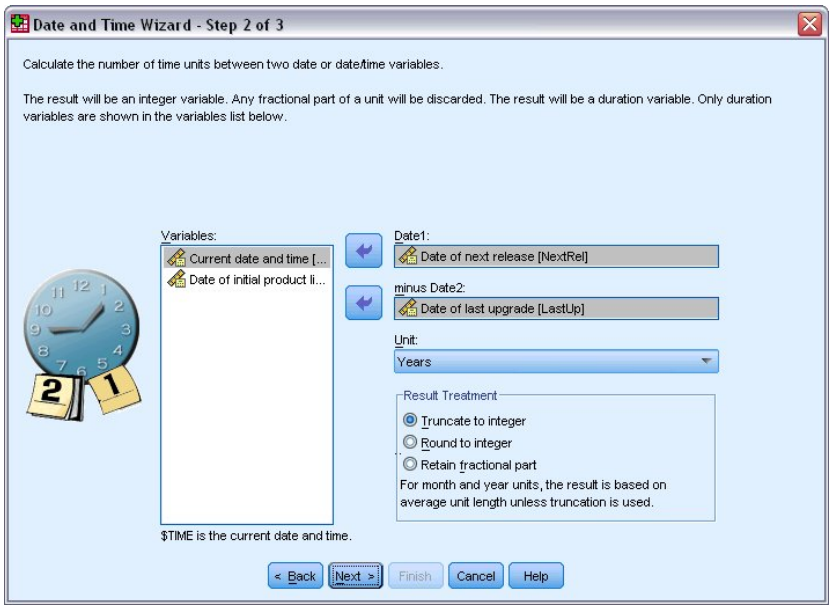


Figure 77. Calculating the length of time between two dates: Step 2

3. In step 2, select *Date of next release* for Date1.
4. Select *Date of last upgrade* for Date2.
5. Select **Years** for the Unit and **Truncate to Integer** for the Result Treatment. (These are the default selections.)
6. Click **Next**.
7. In step 3, enter *YearsLastUp* for the name of the result variable. Result variables cannot have the same name as an existing variable.
8. Enter *Years since last upgrade* as the label for the result variable. Variable labels for result variables are optional.
9. Leave the default selection of **Create the variable now**, and click **Finish** to create the new variable.

The new variable, *YearsLastUp*, displayed in the Data Editor is the integer number of years between the two dates. Fractional parts of a year have been truncated.

Adding a Duration to a Date

You can add or subtract durations, such as 10 days or 12 months, to a date. Continuing with the example of the software company from the previous section, consider determining the date on which each customer's initial tech support contract ends. The data file *upgrade.sav* contains a variable for the number of years of contracted support and a variable for the initial purchase date. You can then determine the end date of the initial support by adding years of support to the purchase date.

To add a duration to a date:

1. Select **Calculate with dates and times** on the introduction screen of the Date and Time Wizard and click **Next**.
2. In step 1, select **Add or subtract a duration from a date** and click **Next**.

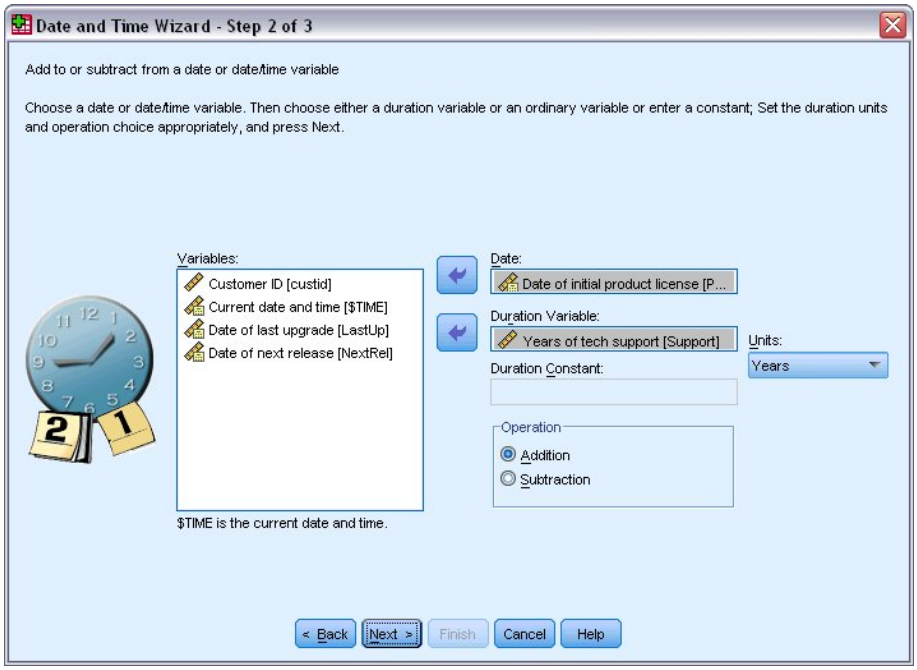


Figure 78. Adding a duration to a date: Step 2

3. Select *Date of initial product license* for Date.
4. In step 2, select *Years of tech support* for the Duration Variable.

   Since *Years of tech support* is simply a numeric variable, you need to indicate the units to use when adding this variable as a duration.
5. Select **Years** from the Units drop-down list.
6. Click **Next**.
7. In step 3, enter *SupEndDate* for the name of the result variable. Result variables cannot have the same name as an existing variable.
8. Enter *End date for support* as the label for the result variable. Variable labels for result variables are optional.
9. Click **Finish** to create the new variable.

The new variable is displayed in the Data Editor.

Chapter 9. Sorting and Selecting Data

Data files are not always organized in the ideal form for your specific needs. To prepare data for analysis, you can select from a wide range of file transformations, including the ability to:

- **Sort data.** You can sort cases based on the value of one or more variables.
- **Select subsets of cases.** You can restrict your analysis to a subset of cases or perform simultaneous analyses on different subsets.

The examples in this chapter use the data file *demo.sav*. See the topic Chapter 10, "Sample Files," on page 83 for more information.

Sorting Data

Sorting cases (sorting rows of the data file) is often useful and sometimes necessary for certain types of analysis.

To reorder the sequence of cases in the data file based on the value of one or more sorting variables:

1. From the menus choose:

   Data > **Sort Cases...**

   The Sort Cases dialog box is displayed.

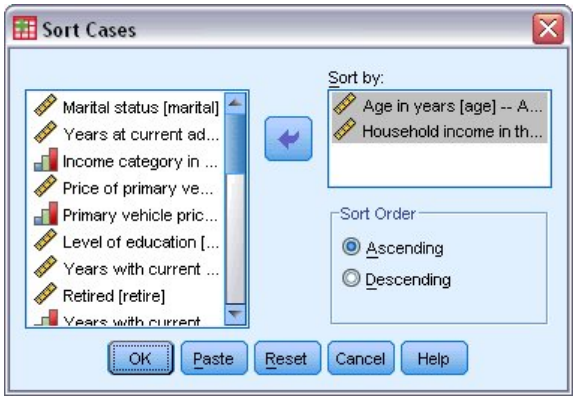


Figure 79. Sort Cases dialog box

2. Add the *Age in years [age]* and *Household income in thousands [income]* variables to the Sort by list.

If you select multiple sort variables, the order in which they appear on the Sort by list determines the order in which cases are sorted. In this example, based on the entries in the Sort by list, cases will be sorted by the value of *Household income in thousands [income]* within categories of *Age in years [age]*. For string variables, uppercase letters precede their lowercase counterparts in sort order (for example, the string value *Yes* comes before *yes* in the sort order).

Split-File Processing

To split your data file into separate groups for analysis:

1. From the menus choose:

   Data > **Split File...**

The Split File dialog box is displayed.

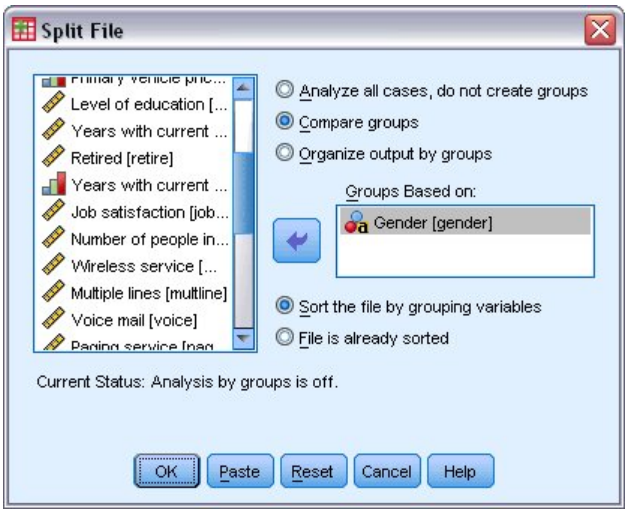


Figure 80. Split File dialog box

2. Select **Compare groups** or **Organize output by groups**. (The examples following these steps show the differences between these two options.)
3. Select *Gender [gender]* to split the file into separate groups for these variables.

You can use numeric, short string, and long string variables as grouping variables. A separate analysis is performed for each subgroup that is defined by the grouping variables. If you select multiple grouping variables, the order in which they appear on the Groups Based on list determines the manner in which cases are grouped.

If you select **Compare groups**, results from all split-file groups will be included in the same table(s), as shown in the following table of summary statistics that is generated by the Frequencies procedure.

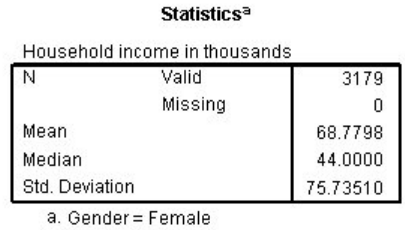
Statistics

Household income in thousands

Female	N	Valid	3179
		Missing	0
	Mean		68.7798
	Median		44.0000
	Std. Deviation		75.73510
Male	N	Valid	3221
		Missing	0
	Mean		70.1608
	Median		45.0000
	Std. Deviation		81.56216

Figure 81. Split-file output with single pivot table

If you select **Organize output by groups** and run the Frequencies procedure, two pivot tables are created: one table for females and one table for males.



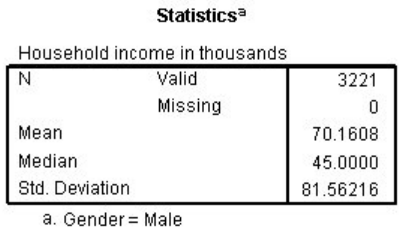
Statistics[a]

Household income in thousands

N	Valid	3179
	Missing	0
Mean		68.7798
Median		44.0000
Std. Deviation		75.73510

a. Gender = Female

Figure 82. Split-file output with pivot table for females



Statistics[a]

Household income in thousands

N	Valid	3221
	Missing	0
Mean		70.1608
Median		45.0000
Std. Deviation		81.56216

a. Gender = Male

Figure 83. Split-file output with pivot table for males

Sorting Cases for Split-File Processing

The Split File procedure creates a new subgroup each time it encounters a different value for one of the grouping variables. Therefore, it is important to sort cases based on the values of the grouping variables before invoking split-file processing.

By default, Split File automatically sorts the data file based on the values of the grouping variables. If the file is already sorted in the proper order, you can save processing time if you select **File is already sorted**.

Turning Split-File Processing On and Off

After you invoke split-file processing, it remains in effect for the rest of the session unless you turn it off.

- **Analyze all cases.** This option turns split-file processing off.
- **Compare groups** and **Organize output by groups.** This option turns split-file processing on.

If split-file processing is in effect, the message **Split File on** appears on the status bar at the bottom of the application window.

Selecting Subsets of Cases

You can restrict your analysis to a specific subgroup based on criteria that include variables and complex expressions. You can also select a random sample of cases. The criteria used to define a subgroup can include:

- Variable values and ranges
- Date and time ranges
- Case (row) numbers
- Arithmetic expressions
- Logical expressions
- Functions

To select a subset of cases for analysis:

1. From the menus choose:
   Data > **Select Cases...**

This opens the Select Cases dialog box.

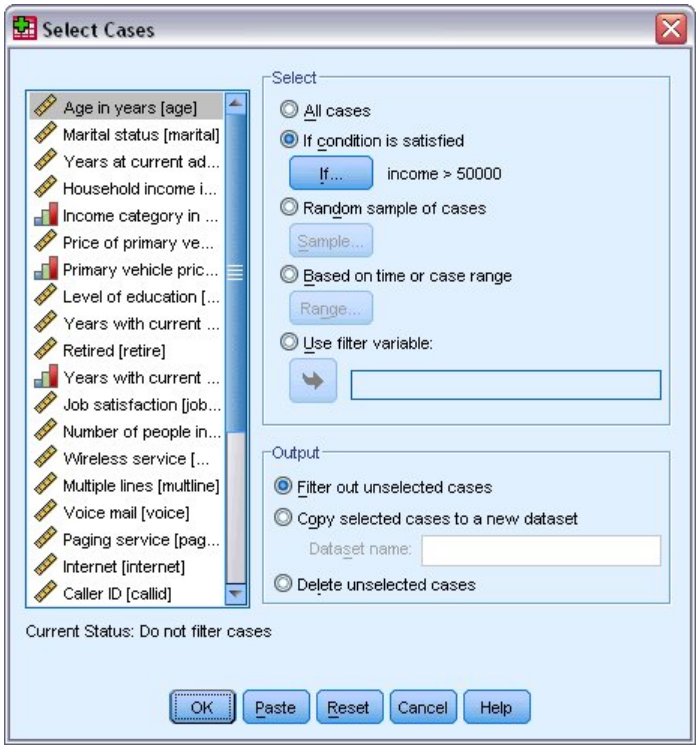


Figure 84. Select Cases dialog box

Selecting Cases Based on Conditional Expressions

To select cases based on a conditional expression:

1. Select **If condition is satisfied** and click **If** in the Select Cases dialog box.

This opens the Select Cases If dialog box.

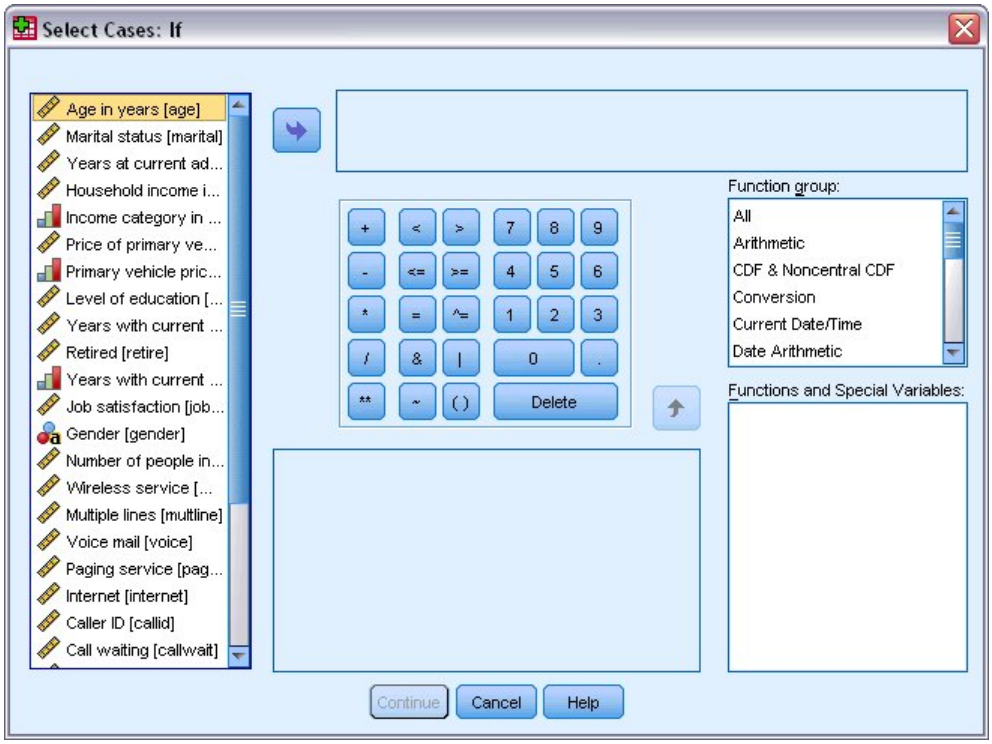


Figure 85. Select Cases If dialog box

The conditional expression can use existing variable names, constants, arithmetic operators, logical operators, relational operators, and functions. You can type and edit the expression in the text box just like text in an output window. You can also use the calculator pad, variable list, and function list to paste elements into the expression. See the topic "Using Conditional Expressions" on page 71 for more information.

Selecting a Random Sample

To obtain a random sample:

1. Select **Random sample of cases** in the Select Cases dialog box.
2. Click **Sample**.

This opens the Select Cases Random Sample dialog box.

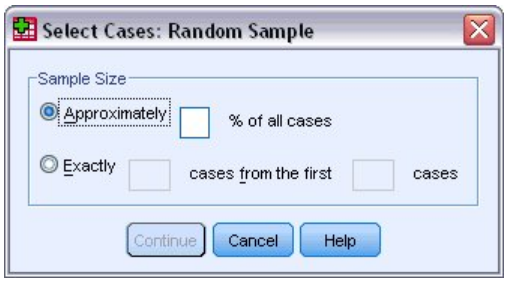


Figure 86. Select Cases Random Sample dialog box

You can select one of the following alternatives for sample size:

- **Approximately.** A user-specified percentage. This option generates a random sample of approximately the specified percentage of cases.

- **Exactly.** A user-specified number of cases. You must also specify the number of cases from which to generate the sample. This second number should be less than or equal to the total number of cases in the data file. If the number exceeds the total number of cases in the data file, the sample will contain proportionally fewer cases than the requested number.

Selecting a Time Range or Case Range

To select a range of cases based on dates, times, or observation (row) numbers:

1. Select **Based on time or case range** and click **Range** in the Select Cases dialog box.

   This opens the Select Cases Range dialog box, in which you can select a range of observation (row) numbers.

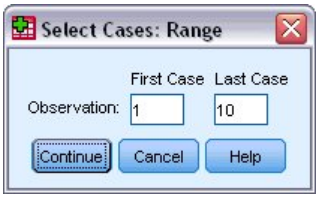


Figure 87. Select Cases Range dialog box

- **First Case.** Enter the starting date and/or time values for the range. If no date variables are defined, enter the starting observation number (row number in the Data Editor, unless Split File is on). If you do not specify a Last Case value, all cases from the starting date/time to the end of the time series are selected.
- **Last Case.** Enter the ending date and/or time values for the range. If no date variables are defined, enter the ending observation number (row number in the Data Editor, unless Split File is on). If you do not specify a First Case value, all cases from the beginning of the time series up to the ending date/time are selected.

For time series data with defined date variables, you can select a range of dates and/or times based on the defined date variables. Each case represents observations at a different time, and the file is sorted in chronological order.

Figure 88. Select Cases Range dialog box (time series)

To generate date variables for time series data:

2. From the menus choose:

   Data > **Define Dates...**

Treatment of Unselected Cases

You can choose one of the following alternatives for the treatment of unselected cases:

- **Filter out unselected cases.** Unselected cases are not included in the analysis but remain in the dataset. You can use the unselected cases later in the session if you turn filtering off. If you select a random sample or if you select cases based on a conditional expression, this generates a variable named *filter_$* with a value of 1 for selected cases and a value of 0 for unselected cases.

- **Copy selected cases to a new dataset.** Selected cases are copied to a new dataset, leaving the original dataset unaffected. Unselected cases are not included in the new dataset and are left in their original state in the original dataset.
- **Delete unselected cases.** Unselected cases are deleted from the dataset. Deleted cases can be recovered only by exiting from the file without saving any changes and then reopening the file. The deletion of cases is permanent if you save the changes to the data file.

Note: If you delete unselected cases and save the file, the cases cannot be recovered.

Case Selection Status

If you have selected a subset of cases but have not discarded unselected cases, unselected cases are marked in the Data Editor with a diagonal line through the row number.

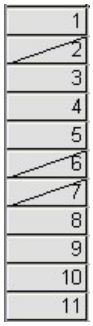


Figure 89. Case selection status

Chapter 10. Sample Files

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

Descriptions

Following are brief descriptions of the sample files used in various examples throughout the documentation.

- **accidents.sav.** This is a hypothetical data file that concerns an insurance company that is studying age and gender risk factors for automobile accidents in a given region. Each case corresponds to a cross-classification of age category and gender.
- **adl.sav.** This is a hypothetical data file that concerns efforts to determine the benefits of a proposed type of therapy for stroke patients. Physicians randomly assigned female stroke patients to one of two groups. The first received the standard physical therapy, and the second received an additional emotional therapy. Three months following the treatments, each patient's abilities to perform common activities of daily life were scored as ordinal variables.
- **advert.sav.** This is a hypothetical data file that concerns a retailer's efforts to examine the relationship between money spent on advertising and the resulting sales. To this end, they have collected past sales figures and the associated advertising costs.
- **aflatoxin.sav.** This is a hypothetical data file that concerns the testing of corn crops for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received 16 samples from each of 8 crop yields and measured the alfatoxin levels in parts per billion (PPB).
- **anorectic.sav.** While working toward a standardized symptomatology of anorectic/bulimic behavior, researchers [1] made a study of 55 adolescents with known eating disorders. Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of 16 symptoms. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations.
- **anticonvulsants.sav.** Medical researchers can use a generalized linear mixed model to determine whether a new anticonvulsant drug can reduce a patient's rate of epileptic seizures. Repeated measurements from the same patient are typically positively correlated so a mixed model with some random effects should be appropriate. The target field, the number of seizures, takes positive integer values, so a generalized linear mixed model with a Poisson distribution and log link may be appropriate.
- **bankloan.sav.** This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.
- **bankloan_binning.sav.** This is a hypothetical data file containing financial and demographic information on 5,000 past customers.

1. Van der Ham, T., J. J. Meulman, D. C. Van Strien, and H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry,* 170, 363-368.

- **bankloan_cs.sav.** This is a hypothetical data file that concerns a bank's efforts to identify characteristics that are indicative of people who are likely to default on loans and then use those characteristics to identify good and bad credit risks.
- **bankloan_cs_noweights.sav.** This is a hypothetical data file that concerns a bank's efforts to identify characteristics that are indicative of people who are likely to default on loans and then use those characteristics to identify good and bad credit risks. The sampling weights are not included in the file.
- **behavior.sav.** In a classic example [2], 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0="extremely appropriate" to 9="extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.
- **behavior_ini.sav.** This data file contains an initial configuration for a two-dimensional solution for *behavior.sav*.
- **brakes.sav.** This is a hypothetical data file that concerns quality control at a factory that produces disc brakes for high-performance automobiles. The data file contains diameter measurements of 16 discs from each of 8 production machines. The target diameter for the brakes is 322 millimeters.
- **breakfast.sav.** In a classic study [3], 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference with 1="most preferred" to 15="least preferred." Their preferences were recorded under six different scenarios, from "Overall preference" to "Snack, with beverage only."
- **breakfast-overall.sav.** This data file contains the breakfast item preferences for the first scenario, "Overall preference," only.
- **broadband_1.sav.** This is a hypothetical data file containing the number of subscribers, by region, to a national broadband service. The data file contains monthly subscriber numbers for 85 regions over a four-year period.
- **broadband_2.sav.** This data file is identical to *broadband_1.sav* but contains data for three additional months.
- **cable_survey.sav.** Executives at a cable provider of television, phone, and internet services want to know more about potential customers. They conduct a survey of 2000 people in their service regions and ask whether they (1) don't have the service; (2) subscribe to the service with other providers; or (3) have the service with the company, for each of the three services. The survey additionally collects some demographic information, such as gender, age category (4 levels), education category (3 levels), income category (3 levels), residence type category (4 levels), years at current address category (3 levels), number of people in the house, and so on.
- **car_insurance_claims.sav.** A dataset presented and analyzed elsewhere [4] concerns damage claims for cars. The average claim amount can be modeled as having a gamma distribution, using an inverse link function to relate the mean of the dependent variable to a linear combination of the policyholder age, vehicle type, and vehicle age. The number of claims filed can be used as a scaling weight.
- **car_sales.sav.** This data file contains hypothetical sales estimates, list prices, and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from *edmunds.com* and manufacturer sites.
- **car_sales_uprepared.sav.** This is a modified version of *car_sales.sav* that does not include any transformed versions of the fields.
- **carpet.sav.** In a popular example [5], a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. Ten consumers rank

2. Price, R. H., and D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology,* 30, 579-586.

3. Green, P. E., and V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

4. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

5. Green, P. E., and Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile.

- **carpet_prefs.sav.** This data file is based on the same example as described for *carpet.sav*, but it contains the actual rankings collected from each of the 10 consumers. The consumers were asked to rank the 22 product profiles from the most to the least preferred. The variables *PREF1* through *PREF22* contain the identifiers of the associated profiles, as defined in *carpet_plan.sav*.
- **catalog.sav.** This data file contains hypothetical monthly sales figures for three products sold by a catalog company. Data for five possible predictor variables are also included.
- **catalog_seasfac.sav.** This data file is the same as *catalog.sav* except for the addition of a set of seasonal factors calculated from the Seasonal Decomposition procedure along with the accompanying date variables.
- **cellular.sav.** This is a hypothetical data file that concerns a cellular phone company's efforts to reduce churn. Churn propensity scores are applied to accounts, ranging from 0 to 100. Accounts scoring 50 or above may be looking to change providers.
- **ceramics.sav.** This is a hypothetical data file that concerns a manufacturer's efforts to determine whether a new premium alloy has a greater heat resistance than a standard alloy. Each case represents a separate test of one of the alloys; the heat at which the bearing failed is recorded.
- **cereal.sav.** This is a hypothetical data file that concerns a poll of 880 people about their breakfast preferences, also noting their age, gender, marital status, and whether or not they have an active lifestyle (based on whether they exercise at least twice a week). Each case represents a separate respondent.
- **clothing_defects.sav.** This is a hypothetical data file that concerns the quality control process at a clothing factory. From each lot produced at the factory, the inspectors take a sample of clothes and count the number of clothes that are unacceptable.
- **coffee.sav.** This data file pertains to perceived images of six iced-coffee brands [6] . For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted AA, BB, CC, DD, EE, and FF to preserve confidentiality.
- **contacts.sav.** This is a hypothetical data file that concerns the contact lists for a group of corporate computer sales representatives. Each contact is categorized by the department of the company in which they work and their company ranks. Also recorded are the amount of the last sale made, the time since the last sale, and the size of the contact's company.
- **credit_card.sav.** A hypothetical study of credit card usage follows each subject's monthly spending on their primary card for two years, with spending broken out by the type of transaction (Grocery, Retail, Entertainment, Travel, and Other). Each record in the dataset corresponds to given month of spending and type of transaction, so the data collected for each subject requires 2 years × 12 months per year × 5 types of transactions = 120 records.
- **creditpromo.sav.** This is a hypothetical data file that concerns a department store's efforts to evaluate the effectiveness of a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months. Half received a standard seasonal ad.
- **cross_sell.sav.** An order-by-mail company has a book club and a CD club. Each month, they make special offers available to club members. The company wants to create a model for the month's total special offer purchases based on total book purchases, CD purchases, and the type of offer given to club members. Two-Stage Least-Squares Regression is appropriate to this situation because money spent on special offers is money not spent on books or CD's; thus, there is a feedback loop between the response and these two predictors.

6. Kennedy, R., C. Riquier, and B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56-70.

- **customer_dbase.sav.** This is a hypothetical data file that concerns a company's efforts to use the information in its data warehouse to make special offers to customers who are most likely to reply. A subset of the customer base was selected at random and given the special offers, and their responses were recorded.
- **customer_information.sav.** A hypothetical data file containing customer mailing information, such as name and address.
- **customer_subset.sav.** A subset of 80 cases from *customer_dbase.sav*.
- **debate.sav.** This is a hypothetical data file that concerns paired responses to a survey from attendees of a political debate before and after the debate. Each case corresponds to a separate respondent.
- **debate_aggregate.sav.** This is a hypothetical data file that aggregates the responses in *debate.sav*. Each case corresponds to a cross-classification of preference before and after the debate.
- **demo.sav.** This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.
- **demo_cs_1.sav.** This is a hypothetical data file that concerns the first step of a company's efforts to compile a database of survey information. Each case corresponds to a different city, and the region, province, district, and city identification are recorded.
- **demo_cs_2.sav.** This is a hypothetical data file that concerns the second step of a company's efforts to compile a database of survey information. Each case corresponds to a different household unit from cities selected in the first step, and the region, province, district, city, subdivision, and unit identification are recorded. The sampling information from the first two stages of the design is also included.
- **demo_cs.sav.** This is a hypothetical data file that contains survey information collected using a complex sampling design. Each case corresponds to a different household unit, and various demographic and sampling information is recorded.
- **diabetes_costs.sav.** This is a hypothetical data file that contains information that is maintained by an insurance company on policy holders who have diabetes. Each case corresponds to a different policy holder.
- **dietstudy.sav.** This hypothetical data file contains the results of a study of the "Stillman diet" [7]. Each case corresponds to a separate subject and records his or her pre- and post-diet weights in pounds and triglyceride levels in mg/100 ml.
- **dmdata.sav.** This is a hypothetical data file that contains demographic and purchasing information for a direct marketing company. *dmdata2.sav* contains information for a subset of contacts that received a test mailing, and *dmdata3.sav* contains information on the remaining contacts who did not receive the test mailing.
- **dvdplayer.sav.** This is a hypothetical data file that concerns the development of a new DVD player. Using a prototype, the marketing team has collected focus group data. Each case corresponds to a separate surveyed user and records some demographic information about them and their responses to questions about the prototype.
- **Employee data.sav.** This is a hypothetical data file that contains employee specific information (education level, employment category, current salary, previous experience, and so on).
- **german_credit.sav.** This data file is taken from the "German credit" dataset in the Repository of Machine Learning Databases [8] at the University of California, Irvine.
- **grocery_1month.sav.** This hypothetical data file is the *grocery_coupons.sav* data file with the weekly purchases "rolled-up" so that each case corresponds to a separate customer. Some of the variables that changed weekly disappear as a result, and the amount spent recorded is now the sum of the amounts spent during the four weeks of the study.

---

7. Rickman, R., N. Mitchell, J. Dingman, and J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228:, 54-58.

8. Blake, C. L., and C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

- **grocery_coupons.sav.** This is a hypothetical data file that contains survey data collected by a grocery store chain interested in the purchasing habits of their customers. Each customer is followed for four weeks, and each case corresponds to a separate customer-week and records information about where and how the customer shops, including how much was spent on groceries during that week.
- **guttman.sav.** Bell [9] presented a table to illustrate possible social groups. Guttman [10] used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups (voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).
- **health_funding.sav.** This is a hypothetical data file that contains data on health care funding (amount per 100 population), disease rates (rate per 10,000 population), and visits to health care providers (rate per 10,000 population). Each case represents a different city.
- **hivassay.sav.** This is a hypothetical data file that concerns the efforts of a pharmaceutical lab to develop a rapid assay for detecting HIV infection. The results of the assay are eight deepening shades of red, with deeper shades indicating greater likelihood of infection. A laboratory trial was conducted on 2,000 blood samples, half of which were infected with HIV and half of which were clean.
- **hourlywagedata.sav.** This is a hypothetical data file that concerns the hourly wages of nurses from office and hospital positions and with varying levels of experience.
- **insurance_claims.sav.** This is a hypothetical data file that concerns an insurance company that wants to build a model for flagging suspicious, potentially fraudulent claims. Each case represents a separate claim.
- **insure.sav.** This is a hypothetical data file that concerns an insurance company that is studying the risk factors that indicate whether a client will have to make a claim on a 10-year term life insurance contract. Each case in the data file represents a pair of contracts, one of which recorded a claim and the other didn't, matched on age and gender.
- **judges.sav.** This is a hypothetical data file that concerns the scores given by trained judges (plus one enthusiast) to 300 gymnastics performances. Each row represents a separate performance; the judges viewed the same performances.
- **kinship_dat.sav.** Rosenberg and Kim [11] set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six “sources” were obtained. Each source corresponds to a 15 x 15 proximity matrix, whose cells are equal to the number of people in a source minus the number of times the objects were partitioned together in that source.
- **kinship_ini.sav.** This data file contains an initial configuration for a three-dimensional solution for *kinship_dat.sav*.
- **kinship_var.sav.** This data file contains independent variables *gender*, *gener*(ation), and *degree* (of separation) that can be used to interpret the dimensions of a solution for *kinship_dat.sav*. Specifically, they can be used to restrict the space of the solution to a linear combination of these variables.
- **marketvalues.sav.** This data file concerns home sales in a new housing development in Algonquin, Ill., during the years from 1999–2000. These sales are a matter of public record.
- **nhis2000_subset.sav.** The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative

9. Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

10. Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika,* 33, 469-506.

11. Rosenberg, S., and M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research,* 10, 489-502.

sample of households. Demographic information and observations about health behaviors and status are obtained for members of each household. This data file contains a subset of information from the 2000 survey. National Center for Health Statistics. National Health Interview Survey, 2000. Public-use data file and documentation. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accessed 2003.

- **ozone.sav.** The data include 330 observations on six meteorological variables for predicting ozone concentration from the remaining variables. Previous researchers [12], [13], among others found nonlinearities among these variables, which hinder standard regression approaches.
- **pain_medication.sav.** This hypothetical data file contains the results of a clinical trial for anti-inflammatory medication for treating chronic arthritic pain. Of particular interest is the time it takes for the drug to take effect and how it compares to an existing medication.
- **patient_los.sav.** This hypothetical data file contains the treatment records of patients who were admitted to the hospital for suspected myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **patlos_sample.sav.** This hypothetical data file contains the treatment records of a sample of patients who received thrombolytics during treatment for myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **poll_cs.sav.** This is a hypothetical data file that concerns pollsters' efforts to determine the level of public support for a bill before the legislature. The cases correspond to registered voters. Each case records the county, township, and neighborhood in which the voter lives.
- **poll_cs_sample.sav.** This hypothetical data file contains a sample of the voters listed in *poll_cs.sav*. The sample was taken according to the design specified in the *poll.csplan* plan file, and this data file records the inclusion probabilities and sample weights. Note, however, that because the sampling plan makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll_jointprob.sav*). The additional variables corresponding to voter demographics and their opinion on the proposed bill were collected and added the data file after the sample as taken.
- **property_assess.sav.** This is a hypothetical data file that concerns a county assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties sold in the county in the past year. Each case in the data file records the township in which the property lies, the assessor who last visited the property, the time since that assessment, the valuation made at that time, and the sale value of the property.
- **property_assess_cs.sav.** This is a hypothetical data file that concerns a state assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties in the state. Each case in the data file records the county, township, and neighborhood in which the property lies, the time since the last assessment, and the valuation made at that time.
- **property_assess_cs_sample.sav.** This hypothetical data file contains a sample of the properties listed in *property_assess_cs.sav*. The sample was taken according to the design specified in the *property_assess.csplan* plan file, and this data file records the inclusion probabilities and sample weights. The additional variable *Current value* was collected and added to the data file after the sample was taken.
- **recidivism.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender and records their demographic information, some details of their first crime, and then the time until their second arrest, if it occurred within two years of the first arrest.
- **recidivism_cs_sample.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender, released from their first arrest during the month of June, 2003, and records their demographic information, some details of their first crime, and the data of their second arrest, if it occurred by the end of June, 2006. Offenders were selected from sampled departments according to the

12. Breiman, L., and J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association,* 80, 580-598.

13. Hastie, T., and R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

sampling plan specified in *recidivism_cs.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*recidivism_cs_jointprob.sav*).

- **rfm_transactions.sav.** A hypothetical data file containing purchase transaction data, including date of purchase, item(s) purchased, and monetary amount of each transaction.
- **salesperformance.sav.** This is a hypothetical data file that concerns the evaluation of two new sales training courses. Sixty employees, divided into three groups, all receive standard training. In addition, group 2 gets technical training; group 3, a hands-on tutorial. Each employee was tested at the end of the training course and their score recorded. Each case in the data file represents a separate trainee and records the group to which they were assigned and the score they received on the exam.
- **satisf.sav.** This is a hypothetical data file that concerns a satisfaction survey conducted by a retail company at 4 store locations. 582 customers were surveyed in all, and each case represents the responses from a single customer.
- **screws.sav.** This data file contains information on the characteristics of screws, bolts, nuts, and tacks [14].
- **shampoo_ph.sav.** This is a hypothetical data file that concerns the quality control at a factory for hair products. At regular time intervals, six separate output batches are measured and their pH recorded. The target range is 4.5–5.5.
- **ships.sav.** A dataset presented and analyzed elsewhere [15] that concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the ship type, construction period, and service period. The aggregate months of service for each cell of the table formed by the cross-classification of factors provides values for the exposure to risk.
- **site.sav.** This is a hypothetical data file that concerns a company's efforts to choose new sites for their expanding business. They have hired two consultants to separately evaluate the sites, who, in addition to an extended report, summarized each site as a "good," "fair," or "poor" prospect.
- **smokers.sav.** This data file is abstracted from the 1998 National Household Survey of Drug Abuse and is a probability sample of American households. (http://dx.doi.org/10.3886/ICPSR02934) Thus, the first step in an analysis of this data file should be to weight the data to reflect population trends.
- **stocks.sav** This hypothetical data file contains stocks prices and volume for one year.
- **stroke_clean.sav.** This hypothetical data file contains the state of a medical database after it has been cleaned using procedures in Statistics Base Edition.
- **stroke_invalid.sav.** This hypothetical data file contains the initial state of a medical database and contains several data entry errors.
- **stroke_survival.** This hypothetical data file concerns survival times for patients exiting a rehabilitation program post-ischemic stroke face a number of challenges. Post-stroke, the occurrence of myocardial infarction, ischemic stroke, or hemorrhagic stroke is noted and the time of the event recorded. The sample is left-truncated because it only includes patients who survived through the end of the rehabilitation program administered post-stroke.
- **stroke_valid.sav.** This hypothetical data file contains the state of a medical database after the values have been checked using the Validate Data procedure. It still contains potentially anomalous cases.
- **survey_sample.sav.** This data file contains survey data, including demographic data and various attitude measures. It is based on a subset of variables from the 1998 NORC General Social Survey, although some data values have been modified and additional fictitious variables have been added for demonstration purposes.
- **tcm_kpi.sav.** This is a hypothetical data file that contains values of weekly key performance indicators for a business. It also contains weekly data for a number of controllable metrics over the same time period.
- **tcm_kpi_upd.sav.** This data file is identical to *tcm_kpi.sav* but contains data for four extra weeks.

---

14. Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

15. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

- **telco.sav.** This is a hypothetical data file that concerns a telecommunications company's efforts to reduce churn in their customer base. Each case corresponds to a separate customer and records various demographic and service usage information.
- **telco_extra.sav.** This data file is similar to the *telco.sav* data file, but the "tenure" and log-transformed customer spending variables have been removed and replaced by standardized log-transformed customer spending variables.
- **telco_missing.sav.** This data file is a subset of the *telco.sav* data file, but some of the demographic data values have been replaced with missing values.
- **testmarket.sav.** This hypothetical data file concerns a fast food chain's plans to add a new item to its menu. There are three possible campaigns for promoting the new product, so the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks. Each case corresponds to a separate location-week.
- **testmarket_1month.sav.** This hypothetical data file is the *testmarket.sav* data file with the weekly sales "rolled-up" so that each case corresponds to a separate location. Some of the variables that changed weekly disappear as a result, and the sales recorded is now the sum of the sales during the four weeks of the study.
- **tree_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_credit.sav.** This is a hypothetical data file containing demographic and bank loan history data.
- **tree_missing_data.sav** This is a hypothetical data file containing demographic and bank loan history data with a large number of missing values.
- **tree_score_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_textdata.sav.** A simple data file with only two variables intended primarily to show the default state of variables prior to assignment of measurement level and value labels.
- **tv-survey.sav.** This is a hypothetical data file that concerns a survey conducted by a TV studio that is considering whether to extend the run of a successful program. 906 respondents were asked whether they would watch the program under various conditions. Each row represents a separate respondent; each column is a separate condition.
- **ulcer_recurrence.sav.** This file contains partial information from a study designed to compare the efficacy of two therapies for preventing the recurrence of ulcers. It provides a good example of interval-censored data and has been presented and analyzed elsewhere [16].
- **ulcer_recurrence_recoded.sav.** This file reorganizes the information in *ulcer_recurrence.sav* to allow you model the event probability for each interval of the study rather than simply the end-of-study event probability. It has been presented and analyzed elsewhere [17].
- **verd1985.sav.** This data file concerns a survey [18]. The responses of 15 subjects to 8 variables were recorded. The variables of interest are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal and *age* is scaled as ordinal; all of the other variables are scaled as single nominal.
- **virus.sav.** This is a hypothetical data file that concerns the efforts of an Internet service provider (ISP) to determine the effects of a virus on its networks. They have tracked the (approximate) percentage of infected e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.
- **wheeze_steubenville.sav.** This is a subset from a longitudinal study of the health effects of air pollution on children [19]. The data contain repeated binary measures of the wheezing status for children

16. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.
17. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.
18. Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.
19. Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, and B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, 366-374.

from Steubenville, Ohio, at ages 7, 8, 9 and 10 years, along with a fixed recording of whether or not the mother was a smoker during the first year of the study.

- **workprog.sav.** This is a hypothetical data file that concerns a government works program that tries to place disadvantaged people into better jobs. A sample of potential program participants were followed, some of whom were randomly selected for enrollment in the program, while others were not. Each case represents a separate program participant.
- **worldsales.sav** This hypothetical data file contains sales revenue by continent and product.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

T

V

W

Printed in USA