IBM SPSS Data Preparation 21

Preface

IBM® SPSS® Statistics is a comprehensive system for analyzing data. The Data Preparation optional add-on module provides the additional analytic techniques described in this manual. The Data Preparation add-on module must be used with the SPSS Statistics Core system and is completely integrated into that system.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit *http://www.ibm.com/spss*.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM Corp. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM Corp. web site at *http://www.ibm.com/support*. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Technical Support for Students

If you're a student using a student, academic or grad pack version of any IBM SPSS software product, please see our special online Solutions for Education (*http://www.ibm.com/spss/rd/students/*) pages for students. If you're a student using a university-supplied copy of the IBM SPSS software, please contact the IBM SPSS product coordinator at your university.

Customer Service

If you have any questions concerning your shipment or account, contact your local office. Please have your serial number ready for identification.

Training Seminars

IBM Corp. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, go to *http://www.ibm.com/software/analytics/spss/training*.

Contents

Part I: User's Guide

8   *Automated Data Preparation*   *83*

9   *Identify Unusual Cases*   *106*

10   *Optimal Binning*   *122*

Appendices

Part I:
User's Guide

Introduction to Data Preparation

As computing systems increase in power, appetites for information grow proportionately, leading to more and more data collection—more cases, more variables, and more data entry errors. These errors are the bane of the predictive model forecasts that are the ultimate goal of data warehousing, so you need to keep the data "clean." However, the amount of data warehoused has grown so far beyond the ability to verify the cases manually that it is vital to implement automated processes for validating data.

The Data Preparation add-on module allows you to identify unusual cases and invalid cases, variables, and data values in your active dataset, and prepare data for modeling.

Usage of Data Preparation Procedures

Your usage of Data Preparation procedures depends on your particular needs. A typical route, after loading your data, is:

- **Metadata preparation.** Review the variables in your data file and determine their valid values, labels, and measurement levels. Identify combinations of variable values that are impossible but commonly miscoded. Define validation rules based on this information. This can be a time-consuming task, but it is well worth the effort if you need to validate data files with similar attributes on a regular basis.

- **Data validation.** Run basic checks and checks against defined validation rules to identify invalid cases, variables, and data values. When invalid data are found, investigate and correct the cause. This may require another step through metadata preparation.

- **Model preparation.** Use automated data preparation to obtain transformations of the original fields that will improve model building. Identify potential statistical outliers that can cause problems for many predictive models. Some outliers are the result of invalid variable values that have not been identified. This may require another step through metadata preparation.

Once your data file is "clean," you are ready to build models from other add-on modules.

Validation Rules

A rule is used to determine whether a case is valid. There are two types of validation rules:

- **Single-variable rules.** Single-variable rules consist of a fixed set of checks that apply to a single variable, such as checks for out-of-range values. For single-variable rules, valid values can be expressed as a range of values or a list of acceptable values.

- **Cross-variable rules.** Cross-variable rules are user-defined rules that can be applied to a single variable or a combination of variables. Cross-variable rules are defined by a logical expression that flags invalid values.

Validation rules are saved to the data dictionary of your data file. This allows you to specify a rule once and then reuse it.

Load Predefined Validation Rules

You can quickly obtain a set of ready-to-use validation rules by loading predefined rules from an external data file included in the installation.

To Load Predefined Validation Rules

▶ From the menus choose:
Data > Validation > Load Predefined Rules...

Figure 2-1
Load Predefined Validation Rules
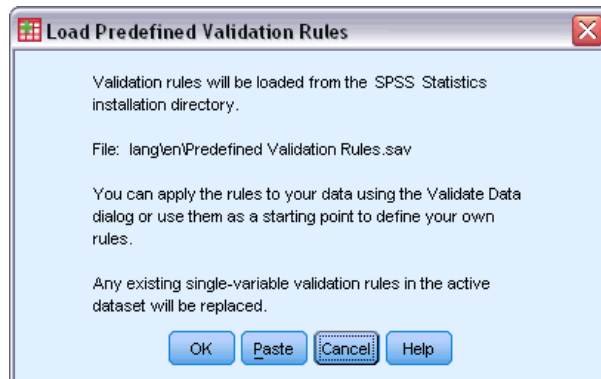
Note that this process deletes any existing single-variable rules in the active dataset.
Alternatively, you can use the Copy Data Properties Wizard to load rules from any data file.

Define Validation Rules

The Define Validation Rules dialog box allows you to create and view single-variable and cross-variable validation rules.

To Create and View Validation Rules

▶ From the menus choose:
Data > Validation > Define Rules...

The dialog box is populated with single-variable and cross-variable validation rules read from the data dictionary. When there are no rules, a new placeholder rule that you can modify to suit your purposes is created automatically.

▶ Select individual rules on the Single-Variable Rules and Cross-Variable Rules tabs to view and modify their properties.

Define Single-Variable Rules

Figure 2-2
Define Validation Rules dialog box, Single-Variable Rules tab
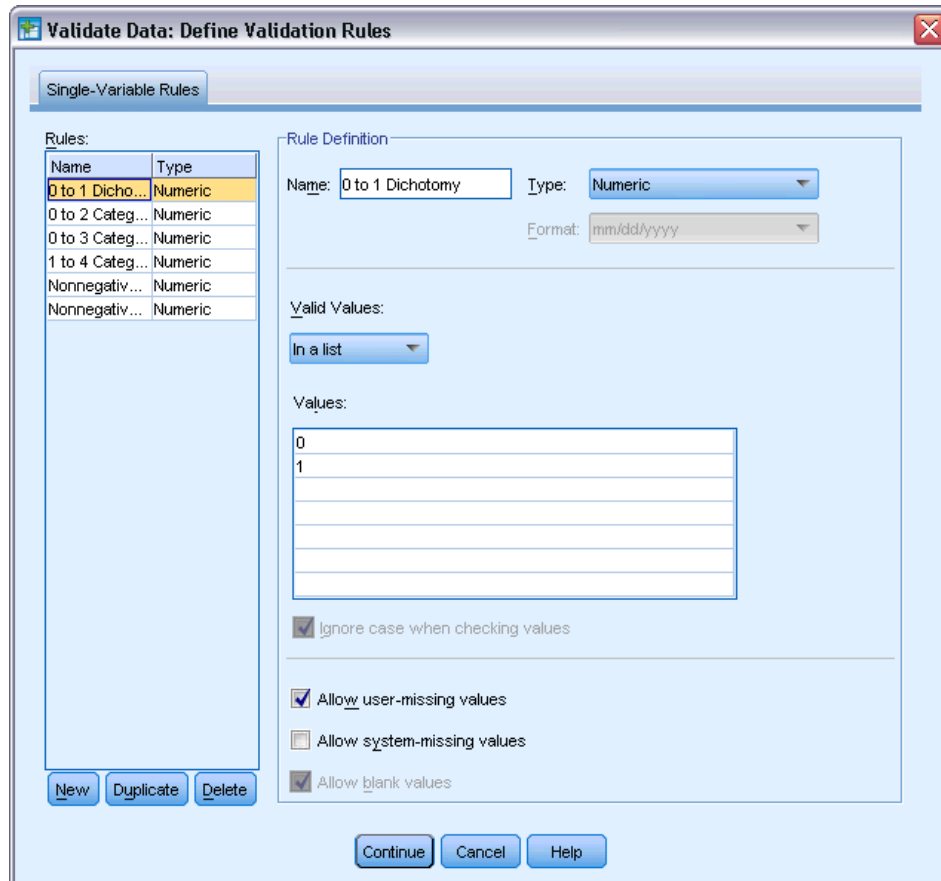
The Single-Variable Rules tab allows you to create, view, and modify single-variable validation rules.

Rules. The list shows single-variable validation rules by name and the type of variable to which the rule can be applied. When the dialog box is opened, it shows rules defined in the data dictionary or, if no rules are currently defined, a placeholder rule called "Single-Variable Rule 1." The following buttons appear below the Rules list:

- **New.** Adds a new entry to the bottom of the Rules list. The rule is selected and assigned the name "SingleVarRule *n*," where *n* is an integer so that the new rule's name is unique among single-variable and cross-variable rules.

- **Duplicate.** Adds a copy of the selected rule to the bottom of the Rules list. The rule name is adjusted so that it is unique among single-variable and cross-variable rules. For example, if you duplicate "SingleVarRule 1," the name of the first duplicate rule would be "Copy of SingleVarRule 1," the second would be "Copy (2) of SingleVarRule 1," and so on.

- **Delete.** Deletes the selected rule.

Rule Definition. These controls allow you to view and set properties for a selected rule.

- **Name.** The name of the rule must be unique among single-variable and cross-variable rules.

- **Type.** This is the type of variable to which the rule can be applied. Select from Numeric, String, and Date.

- **Format.** This allows you to select the date format for rules that can be applied to date variables.

- **Valid Values.** You can specify the valid values either as a range or a list of values.

Range definition controls allow you to specify a valid range. Values outside the range are flagged as invalid.

Figure 2-3
Single-Variable Rules: Range Definition

To specify a range, enter the minimum or maximum values, or both. The check box controls allow you to flag unlabeled and non-integer values within the range.

List definition controls allow you to define a list of valid values. Values not included in the list are flagged as invalid.

Figure 2-4
Single-Variable Rules: List Definition

Enter list values in the grid. The check box determines whether case matters when string data values are checked against the list of acceptable values.

■ **Allow user-missing values.** Controls whether user-missing values are flagged as invalid.

■ **Allow system-missing values.** Controls whether system-missing values are flagged as invalid. This does not apply to string rule types.

■ **Allow blank values.** Controls whether blank (that is, completely empty) string values are flagged as invalid. This does not apply to nonstring rule types.

Define Cross-Variable Rules

Figure 2-5
Define Validation Rules dialog box, Cross-Variable Rules tab
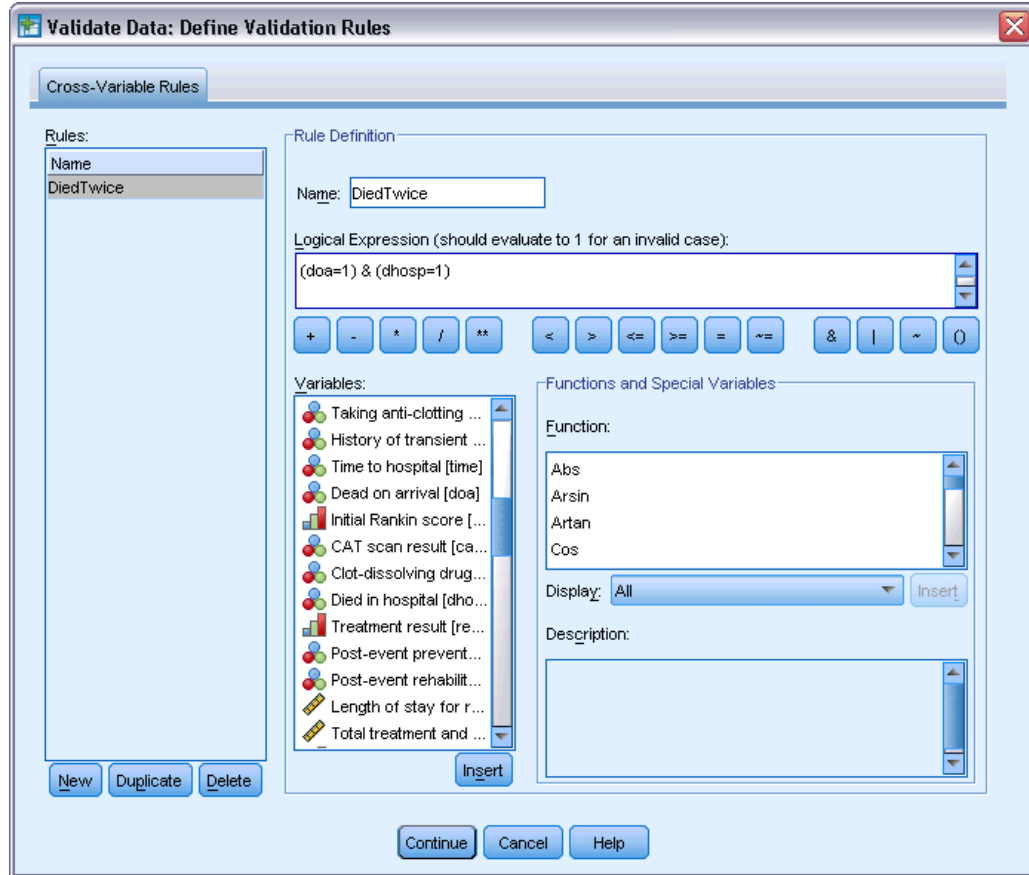
The Cross-Variable Rules tab allows you to create, view, and modify cross-variable validation rules.

Rules. The list shows cross-variable validation rules by name. When the dialog box is opened, it shows a placeholder rule called "CrossVarRule 1." The following buttons appear below the Rules list:

- **New.** Adds a new entry to the bottom of the Rules list. The rule is selected and assigned the name "CrossVarRule *n*," where *n* is an integer so that the new rule's name is unique among single-variable and cross-variable rules.

- **Duplicate.** Adds a copy of the selected rule to the bottom of the Rules list. The rule name is adjusted so that it is unique among single-variable and cross-variable rules. For example, if you duplicate "CrossVarRule 1," the name of the first duplicate rule would be "Copy of CrossVarRule 1," the second would be "Copy (2) of CrossVarRule 1," and so on.

- **Delete.** Deletes the selected rule.

Rule Definition. These controls allow you to view and set properties for a selected rule.

- **Name.** The name of the rule must be unique among single-variable and cross-variable rules.
- **Logical Expression.** This is, in essence, the rule definition. You should code the expression so that invalid cases evaluate to 1.

Building Expressions

▶ To build an expression, either paste components into the Expression field or type directly in the Expression field.

- You can paste functions or commonly used system variables by selecting a group from the Function group list and double-clicking the function or variable in the Functions and Special Variables list (or select the function or variable and click Insert). Fill in any parameters indicated by question marks (applies only to functions). The function group labeled All provides a list of all available functions and system variables. A brief description of the currently selected function or variable is displayed in a reserved area in the dialog box.
- String constants must be enclosed in quotation marks or apostrophes.
- If values contain decimals, a period (.) must be used as the decimal indicator.

Validate Data

The Validate Data dialog box allows you to identify suspicious and invalid cases, variables, and data values in the active dataset.

Example. A data analyst must provide a monthly customer satisfaction report to her client. The data she receives every month needs to be quality checked for incomplete customer IDs, variable values that are out of range, and combinations of variable values that are commonly entered in error. The Validate Data dialog box allows the analyst to specify the variables that uniquely identify customers, define single-variable rules for the valid variable ranges, and define cross-variable rules to catch impossible combinations. The procedure returns a report of the problem cases and variables. Moreover, the data has the same data elements each month, so the analyst is able to apply the rules to the new data file next month.

Statistics. The procedure produces lists of variables, cases, and data values that fail various checks, counts of violations of single-variable and cross-variable rules, and simple descriptive summaries of analysis variables.

Weights. The procedure ignores the weight variable specification and instead treats it as any other analysis variable.

To Validate Data

▶ From the menus choose:
Data > Validation > Validate Data...

Figure 3-1
Validate Data dialog box, Variables tab
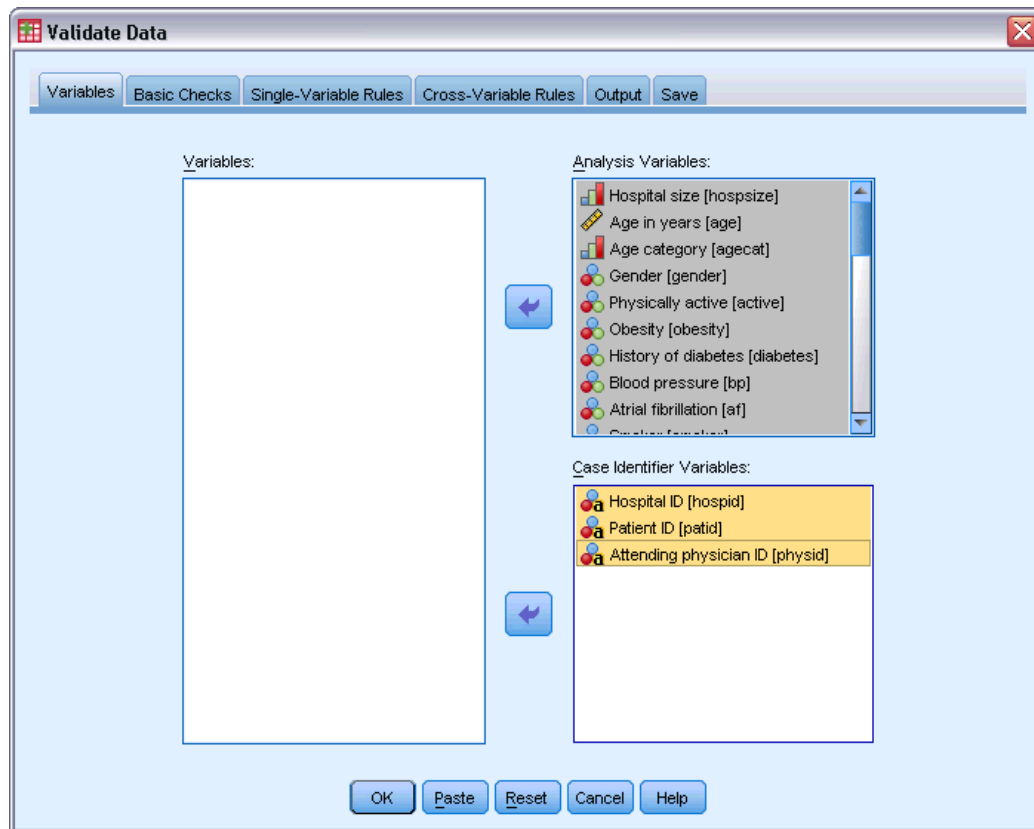
▶ Select one or more analysis variables for validation by basic variable checks or by single-variable validation rules.

Alternatively, you can:

▶ Click the Cross-Variable Rules tab and apply one or more cross-variable rules.

Optionally, you can:

- Select one or more case identification variables to check for duplicate or incomplete IDs. Case ID variables are also used to label casewise output. If two or more case ID variables are specified, the combination of their values is treated as a case identifier.

Fields with Unknown Measurement Level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 3-2
Measurement level alert
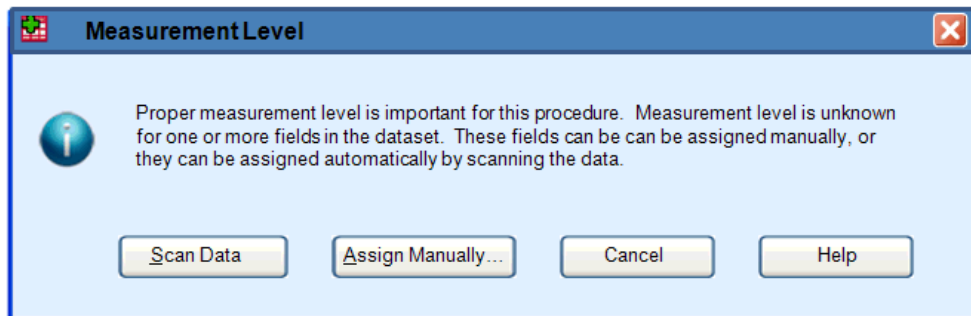
- **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.
- **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Validate Data Basic Checks

Figure 3-3
Validate Data dialog box, Basic Checks tab

The Basic Checks tab allows you to select basic checks for analysis variables, case identifiers, and whole cases.

Analysis Variables. If you selected any analysis variables on the Variables tab, you can select any of the following checks of their validity. The check box allows you to turn the checks on or off.

- **Maximum percentage of missing values.** Reports analysis variables with a percentage of missing values greater than the specified value. The specified value must be a positive number less than or equal to 100.

- **Maximum percentage of cases in a single category.** If any analysis variables are categorical, this option reports categorical analysis variables with a percentage of cases representing a single nonmissing category greater than the specified value. The specified value must be a positive number less than or equal to 100. The percentage is based on cases with nonmissing values of the variable.

- **Maximum percentage of categories with count of 1.** If any analysis variables are categorical, this option reports categorical analysis variables in which the percentage of the variable's categories containing only one case is greater than the specified value. The specified value must be a positive number less than or equal to 100.

■ **Minimum coefficient of variation.** If any analysis variables are scale, this option reports scale analysis variables in which the absolute value of the coefficient of variation is less than the specified value. This option applies only to variables in which the mean is nonzero. The specified value must be a non-negative number. Specifying 0 turns off the coefficient-of-variation check.

■ **Minimum standard deviation.** If any analysis variables are scale, this option reports scale analysis variables whose standard deviation is less than the specified value. The specified value must be a non-negative number. Specifying 0 turns off the standard deviation check.

Case Identifiers. If you selected any case identifier variables on the Variables tab, you can select any of the following checks of their validity.

■ **Flag incomplete IDs.** This option reports cases with incomplete case identifiers. For a particular case, an identifier is considered incomplete if the value of any ID variable is blank or missing.

■ **Flag duplicate IDs.** This option reports cases with duplicate case identifiers. Incomplete identifiers are excluded from the set of possible duplicates.

Flag empty cases. This option reports cases in which all variables are empty or blank. For the purpose of identifying empty cases, you can choose to use all variables in the file (except any ID variables) or only analysis variables defined on the Variables tab.

Validate Data Single-Variable Rules

Figure 3-4
Validate Data dialog box, Single-Variable Rules tab
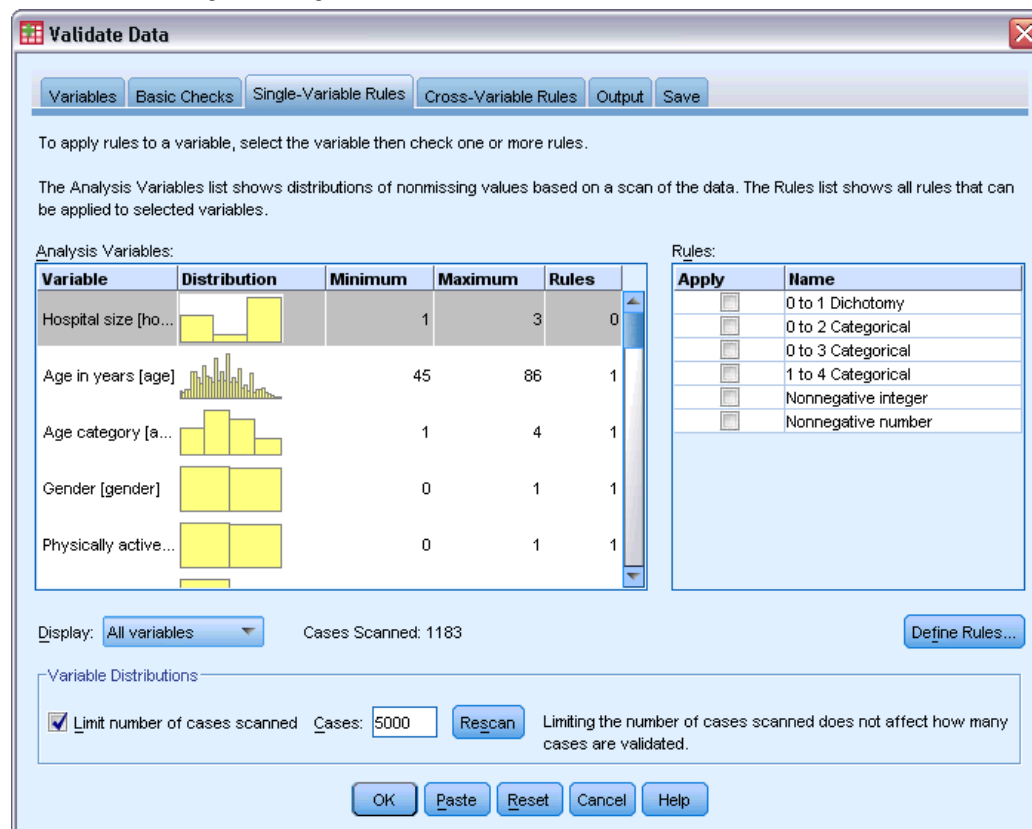
The Single-Variable Rules tab displays available single-variable validation rules and allows you to apply them to analysis variables. To define additional single-variable rules, click Define Rules. For more information, see the topic Define Single-Variable Rules in Chapter 2 on p. 3.

Analysis Variables. The list shows analysis variables, summarizes their distributions, and shows the number of rules applied to each variable. Note that user- and system-missing values are not included in the summaries. The Display drop-down list controls which variables are shown; you can choose from All variables, Numeric variables, String variables, and Date variables.

Rules. To apply rules to analysis variables, select one or more variables and check all rules that you want to apply in the Rules list. The Rules list shows only rules that are appropriate for the selected analysis variables. For example, if numeric analysis variables are selected, only numeric rules are shown; if a string variable is selected, only string rules are shown. If no analysis variables are selected or they have mixed data types, no rules are shown.

Variable Distributions. The distribution summaries shown in the Analysis Variables list can be based on all cases or on a scan of the first *n* cases, as specified in the Cases text box. Clicking Rescan updates the distribution summaries.

Validate Data Cross-Variable Rules

Figure 3-5
Validate Data dialog box, Cross-Variable Rules tab
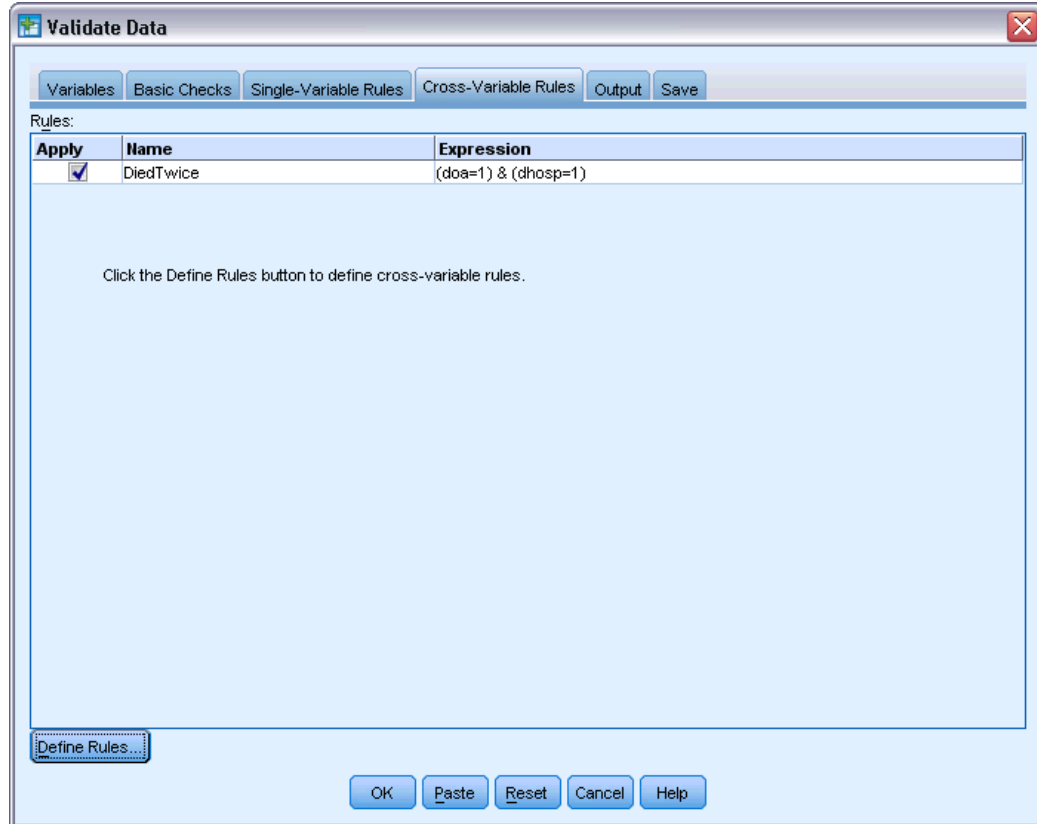
The Cross-Variable Rules tab displays available cross-variable rules and allows you to apply them to your data. To define additional cross-variable rules, click Define Rules. For more information, see the topic Define Cross-Variable Rules in Chapter 2 on p. 6.

Validate Data Output

Figure 3-6
Validate Data dialog box, Output tab
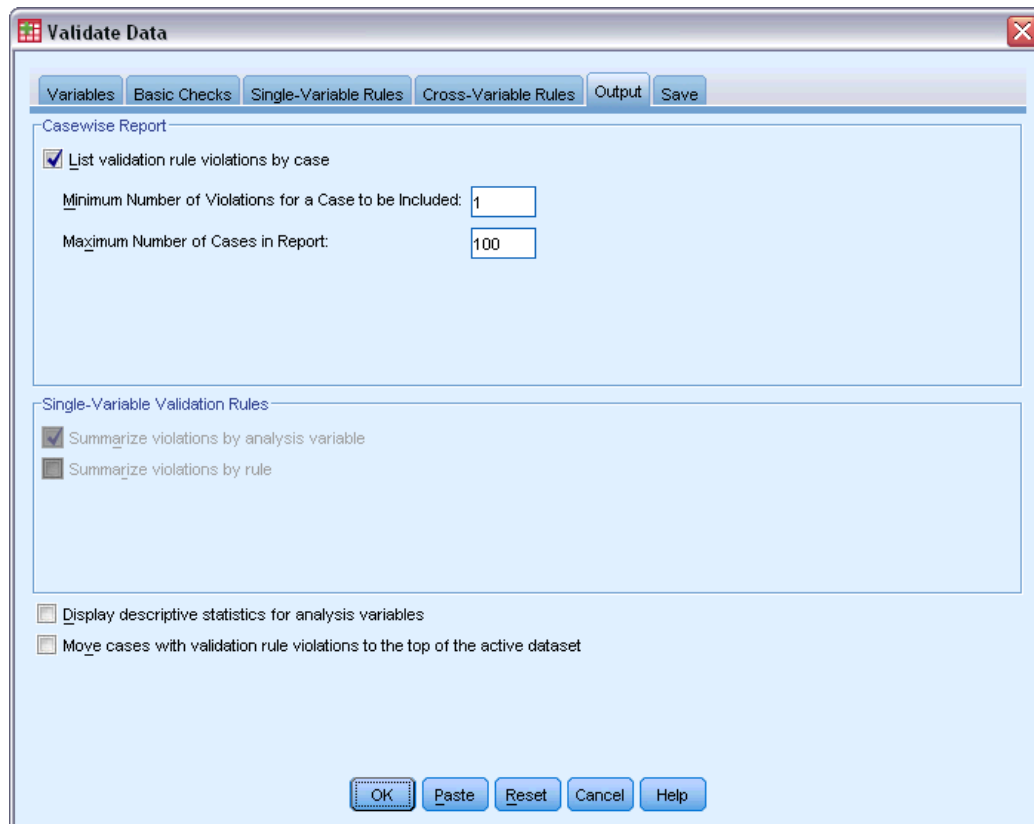
Casewise Report. If you have applied any single-variable or cross-variable validation rules, you can request a report that lists validation rule violations for individual cases.

■ **Minimum Number of Violations.** This option specifies the minimum number of rule violations required for a case to be included in the report. Specify a positive integer.

■ **Maximum Number of Cases.** This option specifies the maximum number of cases included in the case report. Specify a positive integer less than or equal to 1000.

Single-Variable Validation Rules. If you have applied any single-variable validation rules, you can choose how to display the results or whether to display them at all.

■ **Summarize violations by analysis variable.** For each analysis variable, this option shows all single-variable validation rules that were violated and the number of values that violated each rule. It also reports the total number of single-variable rule violations for each variable.

■ **Summarize violations by rule.** For each single-variable validation rule, this option reports variables that violated the rule and the number of invalid values per variable. It also reports the total number of values that violated each rule across variables.

Display descriptive statistics. This option allows you to request descriptive statistics for analysis variables. A frequency table is generated for each categorical variable. A table of summary statistics including the mean, standard deviation, minimum, and maximum is generated for the scale variables.

Move cases with validation rule violations. This option moves cases with single-variable or cross-variable rule violations to the top of the active dataset for easy perusal.

Validate Data Save

Figure 3-7
Validate Data dialog box, Save tab
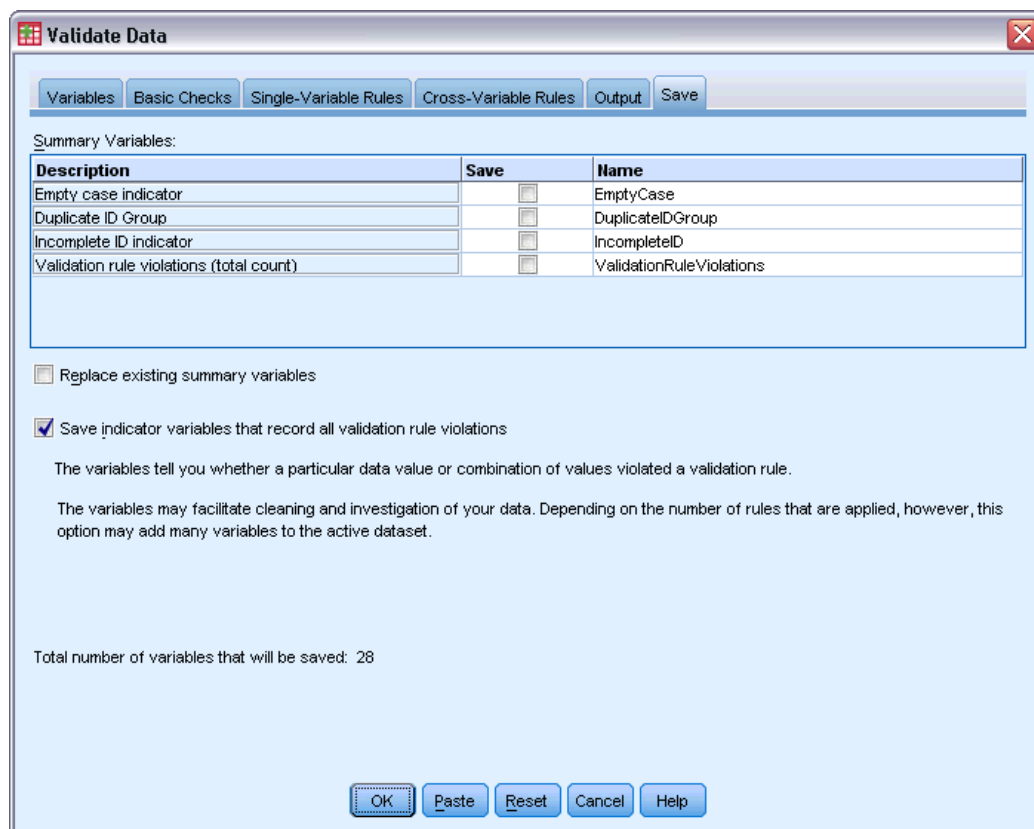
The Save tab allows you to save variables that record rule violations to the active dataset.

Summary Variables. These are individual variables that can be saved. Check a box to save the variable. Default names for the variables are provided; you can edit them.

- **Empty case indicator.** Empty cases are assigned the value 1. All other cases are coded 0. Values of the variable reflect the scope specified on the Basic Checks tab.
- **Duplicate ID Group.** Cases that have the same case identifier (other than cases with incomplete identifiers) are assigned the same group number. Cases with unique or incomplete identifiers are coded 0.

- **Incomplete ID indicator.** Cases with empty or incomplete case identifiers are assigned the value 1. All other cases are coded 0.
- **Validation rule violations.** This is the casewise total count of single-variable and cross-variable validation rule violations.

Replace existing summary variables. Variables saved to the data file must have unique names or replace variables with the same name.

Save indicator variables. This option allows you to save a complete record of validation rule violations. Each variable corresponds to an application of a validation rule and has a value of 1 if the case violates the rule and a value of 0 if it does not.

Automated Data Preparation

Preparing data for analysis is one of the most important steps in any project—and traditionally, one of the most time consuming. Automated Data Preparation (ADP) handles the task for you, analyzing your data and identifying fixes, screening out fields that are problematic or not likely to be useful, deriving new attributes when appropriate, and improving performance through intelligent screening techniques. You can use the algorithm in fully **automatic** fashion, allowing it to choose and apply fixes, or you can use it in **interactive** fashion, previewing the changes before they are made and accept or reject them as desired.

Using ADP enables you to make your data ready for model building quickly and easily, without needing prior knowledge of the statistical concepts involved. Models will tend to build and score more quickly; in addition, using ADP improves the robustness of automated modeling processes.

Note: when ADP prepares a field for analysis, it creates a new field containing the adjustments or transformations, rather than replacing the existing values and properties of the old field. The old field is not used in further analysis; its role is set to None. Also note that any user-missing value information is not transferred to these newly created fields, and any missing values in the new field are system-missing.

Example. An insurance company with limited resources to investigate homeowner's insurance claims wants to build a model for flagging suspicious, potentially fraudulent claims. Before building the model, they will ready the data for modeling using automated data preparation. Since they want to be able to review the proposed transformations before the transformations are applied, they will use automated data preparation in interactive mode. For more information, see the topic Using Automated Data Preparation Interactively in Chapter 8 on p. 83.

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over- and underperforming models, they want to establish a relationship between vehicle sales and vehicle characteristics. They will use automated data preparation to prepare the data for analysis, and build models using the data "before" and "after" preparation to see how the results differ. For more information, see the topic Using Automated Data Preparation Automatically in Chapter 8 on p. 94.

Figure 4-1
Automated Data Preparation Objective tab

Recommends data preparation steps that will speed up model building and improve predictive power. This can include transforming, constructing and selecting features. The target can also be transformed.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings tab that you can further customize, if desired.

- Balance speed & accuracy
- Optimize for speed
- Optimize for accuracy
- Customize analysis

Description

Balanced speed and accuracy adjusts the default setting to transform the data with an emphasis on building models with a balance of speed and accuracy.

What is your objective? Automated data preparation recommends data preparation steps that will affect the speed with which other algorithms can build models and improve the predictive power of those models. This can include transforming, constructing and selecting features. The target can also be transformed. You can specify the model-building priorities that the data preparation process should concentrate on.

■ **Balance speed and accuracy.** This option prepares the data to give equal priority to both the speed with which data are processed by model-building algorithms and the accuracy of the predictions.

■ **Optimize for speed.** This option prepares the data to give priority to the speed with which data are processed by model-building algorithms. When you are working with very large datasets, or are looking for a quick answer, select this option.

■ **Optimize for accuracy.** This option prepares the data to give priority to the accuracy of predictions produced by model-building algorithms.

■ **Custom analysis.** When you want to manually change the algorithm on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with one of the other objectives.

To Obtain Automatic Data Preparation

From the menus choose:
Transform > Prepare Data for Modeling > Automatic...

▶ Click Run.

Optionally, you can:
- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.

To Obtain Interactive Data Preparation

From the menus choose:
Transform > Prepare Data for Modeling > Interactive...

▶ Click Analyze in the toolbar at the top of the dialog.

▶ Click on the Analysis tab and review the suggested data preparation steps.

▶ If satisfied, click Run. Otherwise, click Clear Analysis, change any settings as desired, and click Analyze.

Optionally, you can:
- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.
- Save the suggested data preparation steps to an XML file by clicking Save XML.

Fields Tab

Figure 4-2
Automated Data Preparation Fields tab
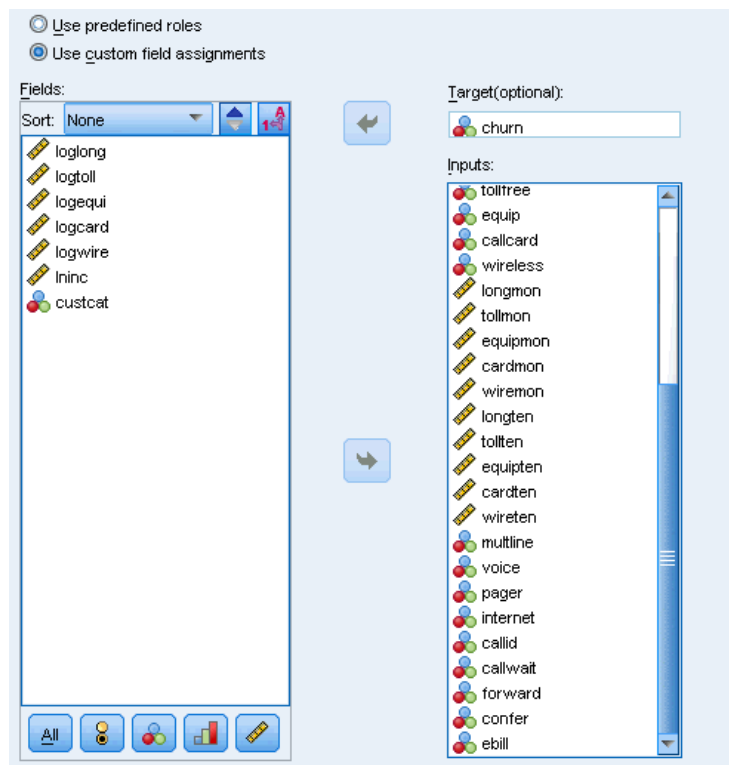
The Fields tab specifies which fields should be prepared for further analysis.

Use predefined roles. This option uses existing field information. If there is a single field with a role as a Target, it will be used as the target; otherwise there will be no target. All fields with a predefined role as an Input will be used as inputs. At least one input field is required.

Use custom field assignments. When you override field roles by moving fields from their default lists, the dialog automatically switches to this option. When making custom field assignments, specify the following fields:

- **Target (optional).** If you plan to build models that require a target, select the target field. This is similar to setting the field role to Target.
- **Inputs.** Select one or more input fields. This is similar to setting the field role to Input.

Settings Tab

The Settings tab comprises several different groups of settings that you can modify to fine-tune how the algorithm processes your data. If you make any changes to the default settings that are incompatible with the other objectives, the Objective tab is automatically updated to select the Customize analysis option.

Prepare Dates & Times

Figure 4-3
Automated Data Preparation Prepare Dates & Times Settings

Many modeling algorithms are unable to directly handle date and time details; these settings enable you to derive new duration data that can be used as model inputs from dates and times in your existing data. The fields containing dates and times must be predefined with date or time storage types. The original date and time fields will not be recommended as model inputs following automated data preparation.

Prepare dates and times for modeling. Deselecting this option disables all other Prepare Dates & Times controls while maintaining the selections.

Compute elapsed time until reference date. This produces the number of years/months/days since a reference date for each variable containing dates.

■ **Reference Date.** Specify the date from which the duration will be calculated with regard to the date information in the input data. Selecting Today's date means that the current system date is always used when ADP is executed. To use a specific date, select Fixed date and enter the required date.

■ **Units for Date Duration.** Specify whether ADP should automatically decide on the date duration unit, or select from Fixed units of Years, Months, or Days.

Compute elapsed time until reference time. This produces the number of hours/minutes/seconds since a reference time for each variable containing times.
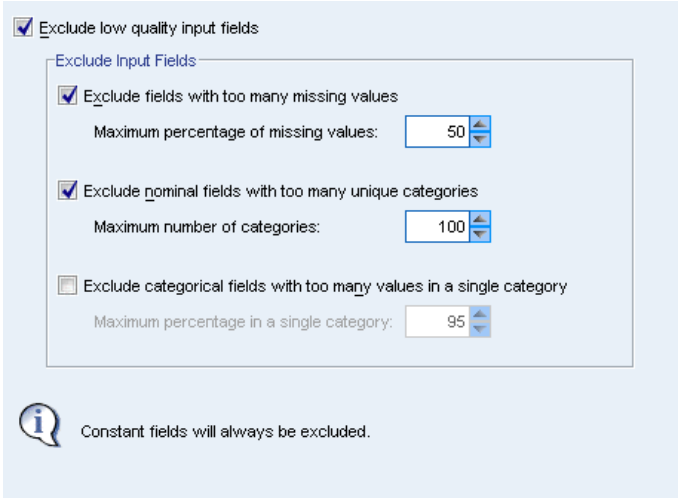
- **Reference Time.** Specify the time from which the duration will be calculated with regard to the time information in the input data. Selecting Current time means that the current system time is always used when ADP is executed. To use a specific time, select Fixed time and enter the required details.

- **Units for Time Duration.** Specify whether ADP should automatically decide on the time duration unit, or select from Fixed units of Hours, Minutes, or Seconds.

Extract Cyclical Time Elements. Use these settings to split a single date or time field into one or more fields. For example if you select all three date checkboxes, the input date field "1954-05-23" is split into three fields: 1954, 5, and 23, each using the suffix defined on the Field Names panel, and the original date field is ignored.

- **Extract from dates.** For any date inputs, specify if you want to extract years, months, days, or any combination.

- **Extract from times.** For any time inputs, specify if you want to extract hours, minutes, seconds, or any combination.

Exclude Fields

Figure 4-4
Automated Data Preparation Exclude Fields Settings

Poor quality data can affect the accuracy of your predictions; therefore, you can specify the acceptable quality level for input features. All fields that are constant or have 100% missing values are automatically excluded.

Exclude low quality input fields. Deselecting this option disables all other Exclude Fields controls while maintaining the selections.
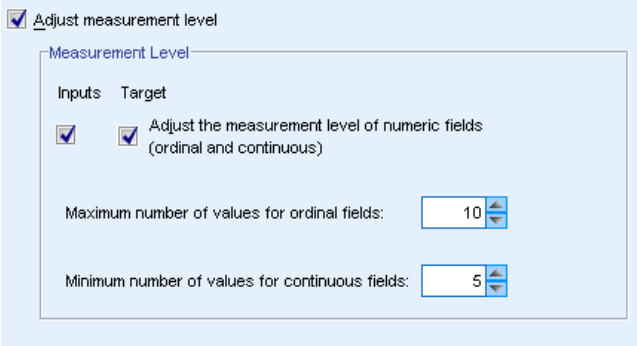
Exclude fields with too many missing values. Fields with more than the specified percentage of missing values are removed from further analysis. Specify a value greater than or equal to 0, which is equivalent to deselecting this option, and less than or equal to 100, though fields with all missing values are automatically excluded. The default is 50.

Exclude nominal fields with too many unique categories. Nominal fields with more than the specified number of categories are removed from further analysis. Specify a positive integer. The default is 100. This is useful for automatically removing fields containing record-unique information from modeling, like ID, address, or name.

Exclude categorical fields with too many values in a single category. Ordinal and nominal fields with a category that contains more than the specified percentage of the records are removed from further analysis. Specify a value greater than or equal to 0, equivalent to deselecting this option, and less than or equal to 100, though constant fields are automatically excluded. The default is 95.

Adjust Measurement

Figure 4-5
Automated Data Preparation Adjust Measurement Settings

Adjust measurement level. Deselecting this option disables all other Adjust Measurement controls while maintaining the selections.

Measurement Level. Specify whether the measurement level of continuous fields with "too few" values can be adjusted to ordinal, and ordinal fields with "too many" values can be adjusted to continuous.

- **Maximum number of values for ordinal fields.** Ordinal fields with more than the specified number of categories are recast as continuous fields. Specify a positive integer. The default is 10. This value must be greater than or equal to the minimum number of values for continuous fields.

- **Minimum number of values for continuous fields.** Continuous fields with less than the specified number of unique values are recast as ordinal fields. Specify a positive integer. The default is 5. This value must be less than or equal to the maximum number of values for ordinal fields.

Improve Data Quality

Figure 4-6
Automated Data Preparation Improve Data Quality Settings

Prepare fields to improve data quality. Deselecting this option disables all other Improve Data Quality controls while maintaining the selections.

Outlier Handling. Specify whether to replace outliers for the inputs and target; if so, specify an outlier cutoff criterion, measured in standard deviations, and a method for replacing outliers. Outliers can be replaced by either trimming (setting to the cutoff value), or by setting them as missing values. Any outliers set to missing values follow the missing value handling settings selected below.

Replace Missing Values. Specify whether to replace missing values of continuous, nominal, or ordinal fields.

Reorder Nominal Fields. Select this to recode the values of nominal (set) fields from smallest (least frequently occurring) to largest (most frequently occurring) category. The new field values start with 0 as the least frequent category. Note that the new field will be numeric even if the original field is a string. For example, if a nominal field's data values are "A", "A", "A", "B", "C", "C", then automated data preparation would recode "B" into 0, "C" into 1, and "A" into 2.

Rescale Fields

Figure 4-7
Automated Data Preparation Rescale Fields Settings

Rescale fields. Deselecting this option disables all other Rescale Fields controls while maintaining the selections.

Analysis Weight. This variable contains analysis (regression or sampling) weights. Analysis weights are used to account for differences in variance across levels of the target field. Select a continuous field.

Continuous Input Fields. This will normalize continuous input fields using a z-score transformation or min/max transformation. Rescaling inputs is especially useful when you select Perform feature construction on the Select and Construct settings.

- **Z-score transformation.** Using the observed mean and standard deviation as population parameter estimates, the fields are standardized and then the *z* scores are mapped to the corresponding values of a normal distribution with the specified Final mean and Final standard deviation. Specify a number for Final mean and a positive number for Final standard deviation. The defaults are 0 and 1, respectively, corresponding to standardized rescaling.

- **Min/max transformation.** Using the observed minimum and maximum as population parameter estimates, the fields are mapped to the corresponding values of a uniform distribution with the specified Minimum and Maximum. Specify numbers with Maximum greater than Minimum.

Continuous Target. This transforms a continuous target using the Box-Cox transformation into a field that has an approximately normal distribution with the specified Final mean and Final standard deviation. Specify a number for Final mean and a positive number for Final standard deviation. The defaults are 0 and 1, respectively.

Note: If a target has been transformed by ADP, subsequent models built using the transformed target score the transformed units. In order to interpret and use the results, you must convert the predicted value back to the original scale. For more information, see the topic Backtransform Scores on p. 44.

Transform Fields

Figure 4-8
Automated Data Preparation Transform Fields Settings

To improve the predictive power of your data, you can transform the input fields.

Transform field for modeling. Deselecting this option disables all other Transform Fields controls while maintaining the selections.

Categorical Input Fields

- **Merge sparse categories to maximize association with target.** Select this to make a more parsimonious model by reducing the number of fields to be processed in association with the target. Similar categories are identified based upon the relationship between the input and the target. Categories that are not significantly different (that is, having a *p*-value greater than the value specified) are merged. Specify a value greater than 0 and less than or equal to 1. If all categories are merged into one, the original and derived versions of the field are excluded from further analysis because they have no value as a predictor.

- **When there is no target, merge sparse categories based on counts.** If the dataset has no target, you can choose to merge sparse categories of ordinal and nominal fields. The equal frequency method is used to merge categories with less than the specified minimum percentage of the total number of records. Specify a value greater than or equal to 0 and less than or equal

to 100. The default is 10. Merging stops when there are not categories with less than the specified minimum percent of cases, or when there are only two categories left.

Continuous Input Fields. If the dataset includes a categorical target, you can bin continuous inputs with strong associations to improve processing performance. Bins are created based upon the properties of "homogeneous subsets", which are identified by the Scheffe method using the specified *p*-value as the alpha for the critical value for determining homogeneous subsets. Specify a value greater than 0 and less than or equal to 1. The default is 0.05. If the binning operation results in a single bin for a particular field, the original and binned versions of the field are excluded because they have no value as a predictor.

Note: Binning in ADP differs from optimal binning. Optimal binning uses entropy information to convert a continuous field to a categorical field; this needs to sort data and store it all in memory. ADP uses homogeneous subsets to bin a continuous field, which means that ADP binning does not need to sort data and does not store all data in memory. The use of the homogeneous subset method to bin a continuous field means that the number of categories after binning is always less than or equal to the number of categories in the target.

Select and Construct

Figure 4-9
Automated Data Preparation Select and Construct Settings

To improve the predictive power of your data, you can construct new fields based on the existing fields.

Perform feature selection. A continuous input is removed from the analysis if the *p*-value for its correlation with the target is greater than the specified *p*-value.

Perform feature construction. Select this option to derive new features from a combination of several existing features. The old features are not used in further analysis. This option only applies to continuous input features where the target is continuous, or where there is no target.

Field Names

Figure 4-10
Automated Data Preparation Name Fields Settings

To easily identify new and transformed features, ADP creates and applies basic new names, prefixes, or suffixes. You can amend these names to be more relevant to your own needs and data.

Transformed and Constructed Fields. Specify the name extensions to be applied to transformed target and input fields.

In addition, specify the prefix name to be applied to any features that are constructed via the Select and Construct settings. The new name is created by attaching a numeric suffix to this prefix root name. The format of the number depends on how many new features are derived, for example:

- 1-9 constructed features will be named: feature1 to feature9.
- 10-99 constructed features will be named: feature01 to feature99.
- 100-999 constructed features will be named: feature001 to feature999, and so on.

This ensures that the constructed features will sort in a sensible order no matter how many there are.

Durations Computed from Dates and Times. Specify the name extensions to be applied to durations computed from both dates and times.
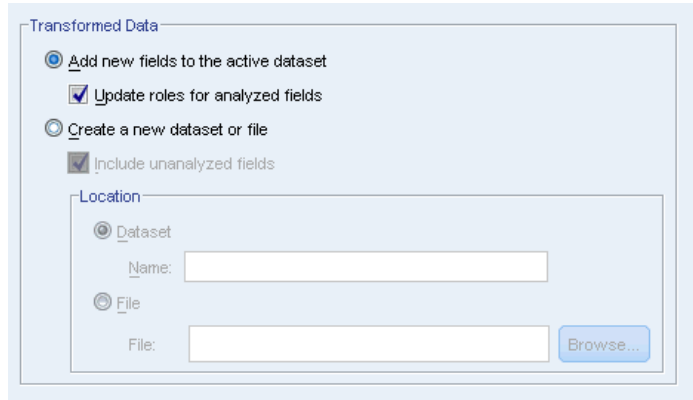
Cyclical Elements Extracted from Dates and Times. Specify the name extensions to be applied to cyclical elements extracted from both dates and times.

Chapter 4

Applying and Saving Transformations

Depending upon whether you are using the Interactive or Automatic Data Preparation dialogs, the settings for applying and saving transformations are slightly different.

Interactive Data Preparation Apply Transformations Settings

Figure 4-11
Interactive Data Preparation Apply Transformations Settings

Transformed Data. These settings specify where to save the transformed data.

- **Add new fields to the active dataset.** Any fields created by automated data preparation are added as new fields to the active dataset. Update roles for analyzed fields will set the role to None for any fields that are excluded from further analysis by automated data preparation.

- **Create a new dataset or file containing the transformed data.** Fields recommended by automated data preparation are added to a new dataset or file. Include unanalyzed fields adds fields in the original dataset that were not specified on the Fields tab to the new dataset. This is useful for transferring fields containing information not used in modeling, like ID or address, or name, into the new dataset.

Automatic Data Preparation Apply and Save Settings

Figure 4-12
Automatic Data Preparation Apply and Save Settings

The Transformed Data group is the same as in Interactive Data Preparation. In Automatic Data preparation, the following additional options are available:

Apply transformations. In the Automatic Data Preparation dialogs, deselecting this option disables all other Apply and Save controls while maintaining the selections.

Save transformations as syntax. This saves the recommended transformations as command syntax to an external file. The Interactive Data Preparation dialog does not have this control because it will paste the transformations as command syntax to the syntax window if you click Paste.

Save transformations as XML. This saves the recommended transformations as XML to an external file, which can be merged with model PMML using TMS MERGE or applied to another dataset using TMS IMPORT. The Interactive Data Preparation dialog does not have this control because it will save the transformations as XML if you click Save XML in the toolbar at the top of the dialog.

Analysis Tab

Note: The Analysis tab is used in the Interactive Data Preparation dialog to allow you to review the recommended transformations. The Automatic Data Preparation dialog does not include this step.

▶ When you are satisfied with the ADP settings, including any changes made on the Objective, Fields, and Settings tabs, click Analyze Data; the algorithm applies the settings to the data inputs and displays the results on the Analysis tab.

The Analysis tab contains both tabular and graphical output that summarizes the processing of your data and displays recommendations as to how the data may be modified or improved for scoring. You can then review and either accept or reject those recommendations.

Figure 4-13
Automated Data Preparation Analysis Tab
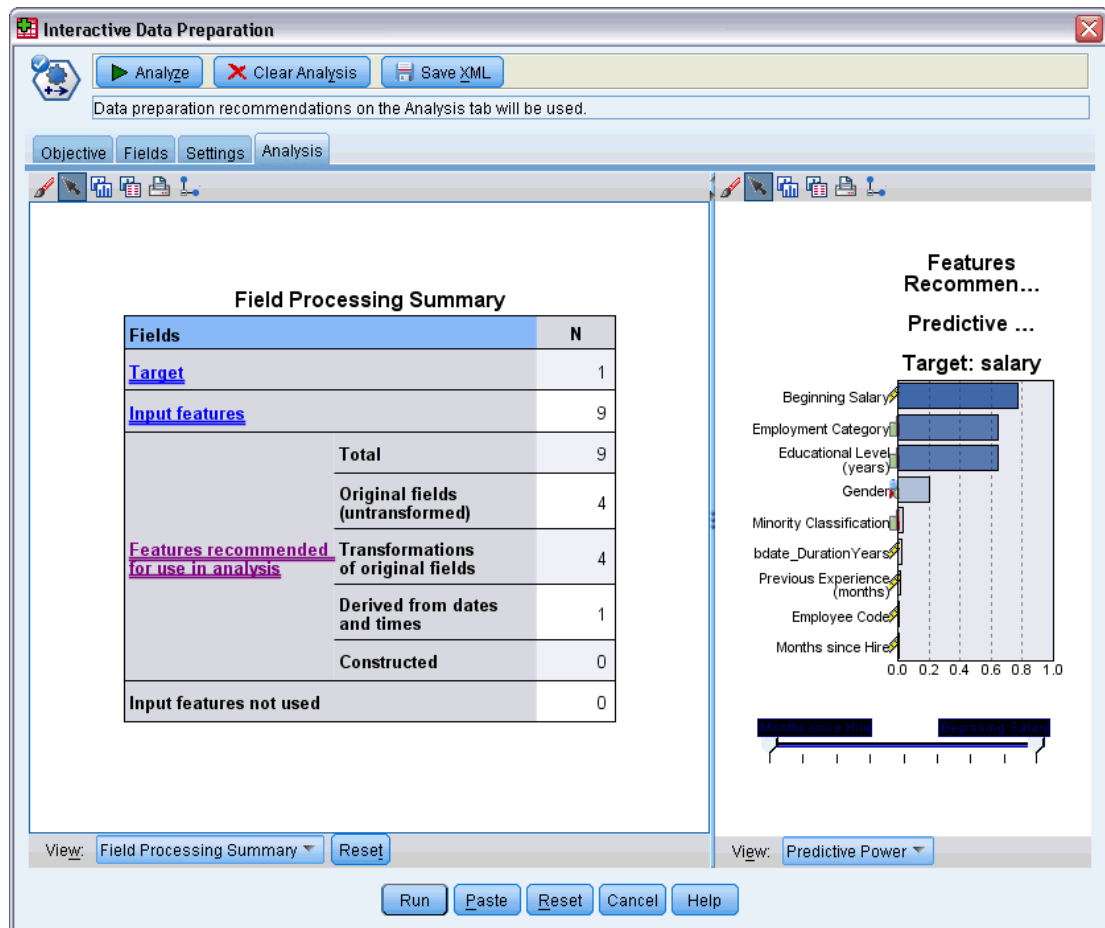
The Analysis tab is made up of two panels, the main view on the left and the linked, or auxiliary, view on the right. There are three main views:

■ Field Processing Summary (the default). For more information, see the topic Field Processing Summary on p. 33.

■ Fields. For more information, see the topic Fields on p. 34.

■ Action Summary. For more information, see the topic Action Summary on p. 36.

There are four linked/auxiliary views:

■ Predictive Power (the default). For more information, see the topic Predictive Power on p. 37.

- Fields Table. For more information, see the topic Fields Table on p. 38.
- Field Details. For more information, see the topic Field Details on p. 39.
- Action Details. For more information, see the topic Action Details on p. 41.

Links between views

Within the main view, underlined text in the tables controls the display in the linked view. Clicking on the text allows you to get details on a particular field, set of fields, or processing step. The link that you last selected is shown in a darker color; this helps you identify the connection between the contents of the two view panels.

Resetting the views

To redisplay the original Analysis recommendations and abandon any changes you have made to the Analysis views, click Reset at the bottom of the main view panel.

Field Processing Summary

Figure 4-14
Field Processing Summary

Field Processing Summary

Fields		N
Target		1
Predictors		9
Predictors recommended for use in analysis	Total	8
	Original fields (untransformed)	1
	Transformations of original fields	7
	Derived from dates and times	0
	Constructed	0
Predictors not used		1

The Field Processing Summary table gives a snapshot of the projected overall impact of processing, including changes to the state of the features and the number of features constructed.

Note that no model is actually built, so there isn't a measure or graph of the change in overall predictive power before and after data preparation; instead, you can display graphs of the predictive power of individual recommended predictors.

The table displays the following information:

■ The number of target fields.

■ The number of original (input) predictors.

■ The predictors recommended for use in analysis and modeling. This includes the total number of fields recommended; the number of original, untransformed, fields recommended; the number of transformed fields recommended (excluding intermediate versions of any field, fields derived from date/time predictors, and constructed predictors); the number of fields recommended that are derived from date/time fields; and the number of constructed predictors recommended.

■ The number of input predictors not recommended for use in any form, whether in their original form, as a derived field, or as input to a constructed predictor.

Where any of the Fields information is underlined, click to display more details in a linked view. Details of the Target, Input features, and Input features not used are shown in the Fields Table linked view. For more information, see the topic Fields Table on p. 38. Features recommended for use in analysis are displayed in the Predictive Power linked view. For more information, see the topic Predictive Power on p. 37.

Fields

Figure 4-15
Fields

The Fields main view displays the processed fields and whether ADP recommends using them in downstream models. You can override the recommendation for any field; for example, to exclude constructed features or include features that ADP recommends excluding. If a field has been transformed, you can decide whether to accept the suggested transformation or use the original version.

The Fields view consists of two tables, one for the target and one for predictors that were either processed or created.

Target table

The Target table is only shown if a target is defined in the data.

The table contains two columns:

- **Name.** This is the name or label of the target field; the original name is always used, even if the field has been transformed.

- **Measurement Level.** This displays the icon representing the measurement level; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.

  If the target has been transformed the Measurement Level column reflects the final transformed version. *Note*: you cannot turn off transformations for the target.

Predictors table

The Predictors table is always shown. Each row of the table represents a field. By default the rows are sorted in descending order of predictive power.

For ordinary features, the original name is always used as the row name. Both original and derived versions of date/time fields appear in the table (in separate rows); the table also includes constructed predictors.

Note that transformed versions of fields shown in the table always represent the final versions.

By default only recommended fields are shown in the Predictors table. To display the remaining fields, select the Include nonrecommended fields in table box above the table; these fields are then displayed at the bottom of the table.

The table contains the following columns:

- **Version to Use.** This displays a drop-down list that controls whether a field will be used downstream and whether to use the suggested transformations. By default, the drop-down list reflects the recommendations.

  For ordinary predictors that have been transformed the drop-down list has three choices: Transformed, Original, and Do not use.

  For untransformed ordinary predictors the choices are: Original and Do not use.

  For derived date/time fields and constructed predictors the choices are: Transformed and Do not use.

For original date fields the drop-down list is disabled and set to Do not use.

Note: For predictors with both original and transformed versions, changing between Original and Transformed versions automatically updates the Measurement Level and Predictive Power settings for those features.

■ **Name.** Each field's name is a link. Click on a name to display more information about the field in the linked view. For more information, see the topic Field Details on p. 39.

■ **Measurement Level.** This displays the icon representing the data type; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.

■ **Predictive Power.** Predictive power is displayed only for fields that ADP recommends. This column is not displayed if there is no target defined. Predictive power ranges from 0 to 1, with larger values indicating "better" predictors. In general, predictive power is useful for comparing predictors within an ADP analysis, but predictive power values should not be compared across analyses.

Action Summary

Figure 4-16
Action Summary

Action Summary

Action
Text Fields
Date and Time Predictors
Predictor Screening
Check Measurement Level
Outliers
Missing Values
Target
Categorical Predictors
Continuous Predictors

For each action taken by automated data preparation, input predictors are transformed and/or filtered out; fields that survive one action are used in the next. The fields that survive through to the last step are then recommended for use in modeling, whilst inputs to transformed and constructed predictors are filtered out.

The Action Summary is a simple table that lists the processing actions taken by ADP. Where any Action is underlined, click to display more details in a linked view about the actions taken. For more information, see the topic Action Details on p. 41.

Note: Only the original and final transformed versions of each field are shown, not any intermediate versions that were used during analysis.

Predictive Power

Figure 4-17
Predictive Power

Displayed by default when the analysis is first run, or when you select Predictors recommended for use in analysis in the Field Processing Summary main view, the chart displays the predictive power of recommended predictors. Fields are sorted by predictive power, with the field with the highest value appearing at the top.

For transformed versions of ordinary predictors, the field name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: _transformed.

Measurement level icons are displayed after the individual field names.

The predictive power of each recommended predictor is computed from either a linear regression or naïve Bayes model, depending upon whether the target is continuous or categorical.

Fields Table

Figure 4-18
Fields Table

Predictors

Name	Measurement Level
ID	Continuous
GENDER	Nominal
BDATE	Continuous
EDUC	Ordinal
JOBCAT	Ordinal
SALBEGIN	Continuous
JOBTIME	Continuous
PREVEXP	Continuous
MINORITY	Ordinal

Displayed when you click Target, Predictors, or Predictors not used in the Field Processing Summary main view, the Fields Table view displays a simple table listing the relevant features.

The table contains two columns:

■ **Name.** The predictor name.

For targets, the original name or label of the field is used, even if the target has been transformed.

For transformed versions of ordinary predictors, the name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: *_transformed.*

For fields derived from dates and times, the name of the final transformed version is used; for example: *bdate_years*.

For constructed predictors, the name of the constructed predictor is used; for example: *Predictor1*.

■ **Measurement Level.** This displays the icon representing the data type.

For the Target, the Measurement Level always reflects the transformed version (if the target has been transformed); for example, changed from ordinal (ordered set) to continuous (range, scale), or vice versa.

Field Details

Figure 4-19
Field Details

Details for Beginning Salary

Original	Transformed

Distribution

Original: Skewness=2.85, SD=7,870.64

Transformed: Skewness=2.85, SD=1.00

Missing Values

Original: 0.0%

Transformed: 0.0%

Predictive Power

Processing

Action	Details
Continuous Predictors	Transform to standard units (Mean=0, SD=1)

Name of Transformed Field: SALBEGIN_transformed
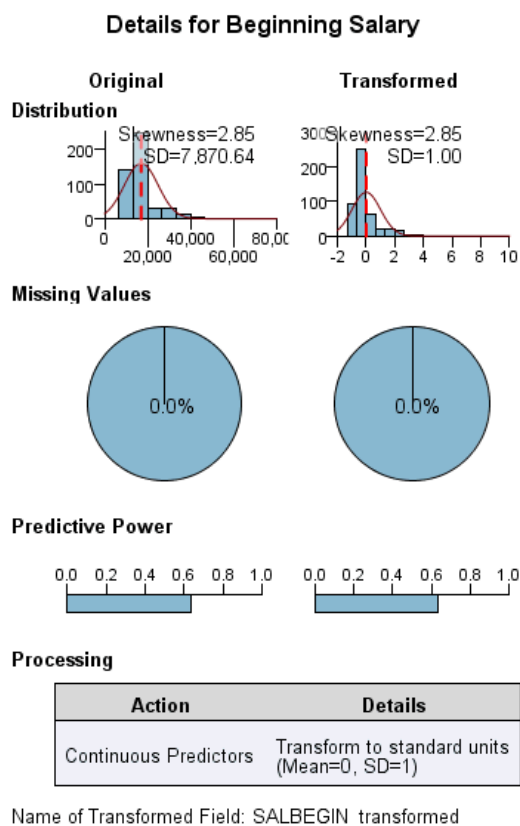
Displayed when you click any Name in the Fields main view, the Field Details view contains distribution, missing values, and predictive power charts (if applicable) for the selected field. In addition, the processing history for the field and the name of the transformed field are also shown (if applicable)

For each chart set, two versions are shown side by side to compare the field with and without transformations applied; if a transformed version of the field does not exist, a chart is shown for the original version only. For derived date or time fields and constructed predictors, the charts are only shown for the new predictor.

Note: If a field is excluded due to having too many categories only the processing history is shown.

Distribution Chart

Continuous field distribution is shown as a histogram, with a normal curve overlaid, and a vertical reference line for the mean value; categorical fields are displayed as a bar chart.

Histograms are labeled to show standard deviation and skewness; however, skewness is not displayed if the number of values is 2 or fewer or the variance of the original field is less than 10-20.

Hover the mouse over the chart to display either the mean for histograms, or the count and percentage of the total number of records for categories in bar charts.

Missing Value Chart

Pie charts compare the percentage of missing values with and without transformations applied; the chart labels show the percentage.

If ADP carried out missing value handling, the post-transformation pie chart also includes the replacement value as a label — that is, the value used in place of missing values.

Hover the mouse over the chart to display the missing value count and percentage of the total number of records.

Predictive Power Chart

For recommended fields, bar charts display the predictive power before and after transformation. If the target has been transformed, the calculated predictive power is in respect to the transformed target.

Note: Predictive power charts are not shown if no target is defined, or if the target is clicked in the main view panel.

Hover the mouse over the chart to display the predictive power value.

Processing History Table

The table shows how the transformed version of a field was derived. Actions taken by ADP are listed in the order in which they were carried out; however, for certain steps multiple actions may have been carried out for a particular field.

Note: This table is not shown for fields that have not been transformed.

The information in the table is broken down into two or three columns:

- **Action.** The name of the action. For example, Continuous Predictors. For more information, see the topic Action Details on p. 41.
- **Details.** The list of processing carried out. For example, Transform to standard units.
- **Function.** Only shown only for constructed predictors, this displays the linear combination of input fields, for example, .06*age + 1.21*height.

Action Details

Figure 4-20
ADP Analysis - Action Details

Step 9: Continuous Predictors

Transformation	Number of Predictors	Criteria	
		Mean	SD
Transform to standard units	5	0	1

Predictor Space Construction	N
Predictors constructed	0
Predictors excluded due to low association with target	1
Predictors excluded because they were constant after binning	0

Displayed when you select any underlined Action in the Action Summary main view, the Action Details linked view displays both action-specific and common information for each processing step that was carried out; the action-specific details are displayed first.

For each action, the description is used as the title at the top of the linked view. The action-specific details are displayed below the title, and may include details of the number of derived predictors, fields recast, target transformations, categories merged or reordered, and predictors constructed or excluded.

As each action is processed, the number of predictors used in the processing may change, for example as predictors are excluded or merged.

Note: If an action was turned off, or no target was specified, an error message is displayed in place of the action details when the action is clicked in the Action Summary main view.

There are nine possible actions; however, not all are necessarily active for every analysis.

Text Fields Table

The table displays the number of:
- Predictors excluded from analysis.

Date and Time Predictors Table

The table displays the number of:
- Durations derived from date and time predictors.
- Date and time elements.
- Derived date and time predictors, in total.

The reference date or time is displayed as a footnote if any date durations were calculated.

Predictor Screening Table

The table displays the number of the following predictors excluded from processing:
- Constants.
- Predictors with too many missing values.
- Predictors with too many cases in a single category.
- Nominal fields (sets) with too many categories.
- Predictors screened out, in total.

Check Measurement Level Table

The table displays the numbers of fields recast, broken down into the following:
- Ordinal fields (ordered sets) recast as continuous fields.
- Continuous fields recast as ordinal fields.
- Total number recast.

If no input fields (target or predictors) were continuous or ordinal, this is shown as a footnote.

Outliers Table

The table displays counts of how any outliers were handled.
- Either the number of continuous fields for which outliers were found and trimmed, or the number of continuous fields for which outliers were found and set to missing, depending on your settings in the Prepare Inputs & Target panel on the Settings tab.
- The number of continuous fields excluded because they were constant, after outlier handling.

One footnote shows the outlier cutoff value; while another footnote is shown if no input fields (target or predictors) were continuous.

Missing Values Table

The table displays the numbers of fields that had missing values replaced, broken down into:
- Target. This row is not shown if no target is specified.

- Predictors. This is further broken down into the number of nominal (set), ordinal (ordered set), and continuous.
- The total number of missing values replaced.

Target Table

The table displays whether the target was transformed, shown as:

- Box-Cox transformation to normality. This is further broken down into columns that show the specified criteria (mean and standard deviation) and Lambda.
- Target categories reordered to improve stability.

Categorical Predictors Table

The table displays the number of categorical predictors:

- Whose categories were reordered from lowest to highest to improve stability.
- Whose categories were merged to maximize association with the target.
- Whose categories were merged to handle sparse categories.
- Excluded due to low association with the target.
- Excluded because they were constant after merging.

A footnote is shown if there were no categorical predictors.

Continuous Predictors Table

There are two tables. The first displays one of the following number of transformations:

- Predictor values transformed to standard units. In addition, this shows the number of predictors transformed, the specified mean, and the standard deviation.
- Predictor values mapped to a common range. In addition, this shows the number of predictors transformed using a min-max transformation, as well as the specified minimum and maximum values.
- Predictor values binned and the number of predictors binned.

The second table displays the predictor space construction details, shown as the number of predictors:

- Constructed.
- Excluded due to a low association with the target.
- Excluded because they were constant after binning.
- Excluded because they were constant after construction.

A footnote is shown if no continuous predictors were input.

Backtransform Scores

If a target has been transformed by ADP, subsequent models built using the transformed target score the transformed units. In order to interpret and use the results, you must convert the predicted value back to the original scale.

Figure 4-21
Backtransform Scores
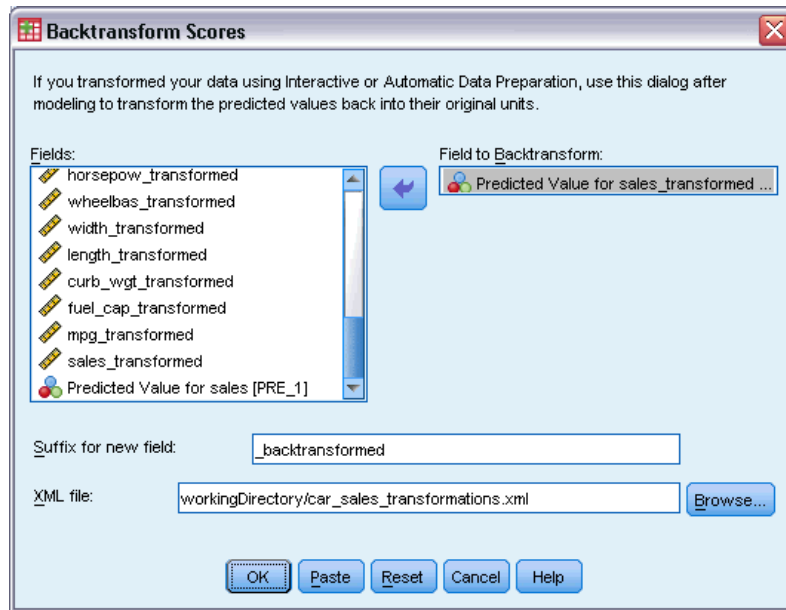
To backtransform scores, from the menus choose:
Transform > Prepare Data for Modeling > Backtransform Scores...

► Select a field to backtransform. This field should contain model-predicted values of the transformed target.

► Specify a suffix for the new field. This new field will contain model-predicted values in the original scale of the untransformed target.

► Specify the location of the XML file containing the ADP transformations. This should be a file saved from the Interactive or Automatic Data Preparation dialogs. For more information, see the topic Applying and Saving Transformations on p. 30.

Identify Unusual Cases

The Anomaly Detection procedure searches for unusual cases based on deviations from the norms of their cluster groups. The procedure is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm is designed for generic anomaly detection; that is, the definition of an anomalous case is not specific to any particular application, such as detection of unusual payment patterns in the healthcare industry or detection of money laundering in the finance industry, in which the definition of an anomaly can be well-defined.

Example. A data analyst hired to build predictive models for stroke treatment outcomes is concerned about data quality because such models can be sensitive to unusual observations. Some of these outlying observations represent truly unique cases and are thus unsuitable for prediction, while other observations are caused by data entry errors in which the values are technically "correct" and thus cannot be caught by data validation procedures. The Identify Unusual Cases procedure finds and reports these outliers so that the analyst can decide how to handle them.

Statistics. The procedure produces peer groups, peer group norms for continuous and categorical variables, anomaly indices based on deviations from peer group norms, and variable impact values for variables that most contribute to a case being considered unusual.

Data Considerations

Data. This procedure works with both continuous and categorical variables. Each row represents a distinct observation, and each column represents a distinct variable upon which the peer groups are based. A case identification variable can be available in the data file for marking output, but it will not be used in the analysis. Missing values are allowed. The weight variable, if specified, is ignored.

The detection model can be applied to a new test data file. The elements of the test data must be the same as the elements of the training data. And, depending on the algorithm settings, the missing value handling that is used to create the model may be applied to the test data file prior to scoring.

Case order. Note that the solution may depend on the order of cases. To minimize order effects, randomly order the cases. To verify the stability of a given solution, you may want to obtain several different solutions with cases sorted in different random orders. In situations with extremely large file sizes, multiple runs can be performed with a sample of cases sorted in different random orders.

Assumptions. The algorithm assumes that all variables are nonconstant and independent and that no case has missing values for any of the input variables. Each continuous variable is assumed to have a normal (Gaussian) distribution, and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but be aware of how well these assumptions are met.

45

To Identify Unusual Cases

▶ From the menus choose:

Data > Identify Unusual Cases...

Figure 5-1
Identify Unusual Cases dialog box, Variables tab
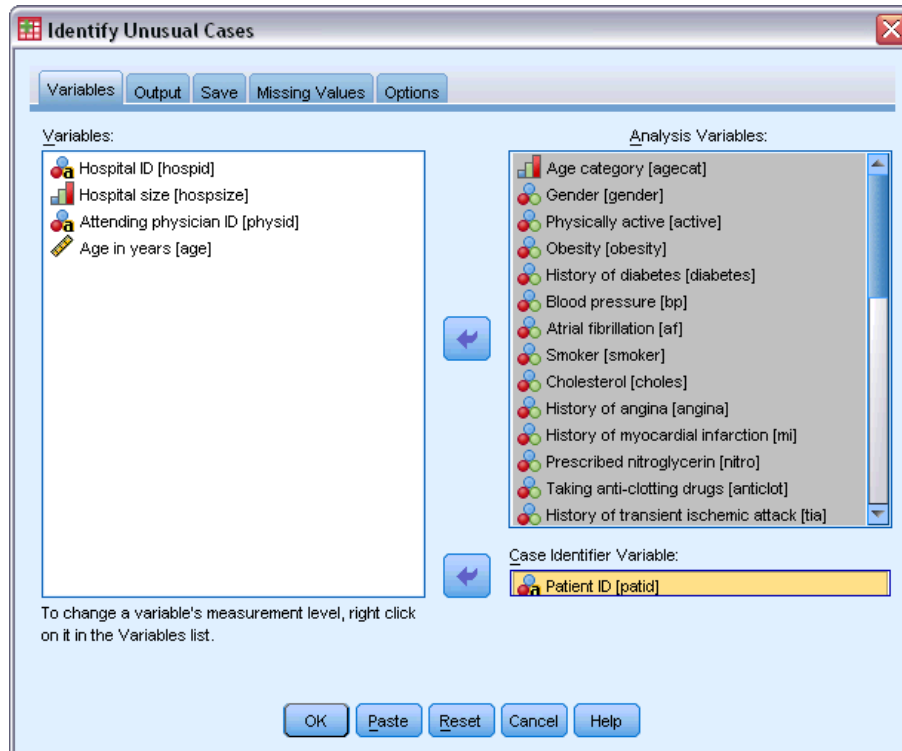
▶ Select at least one analysis variable.

▶ Optionally, choose a case identifier variable to use in labeling output.

Fields with Unknown Measurement Level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 5-2
Measurement level alert
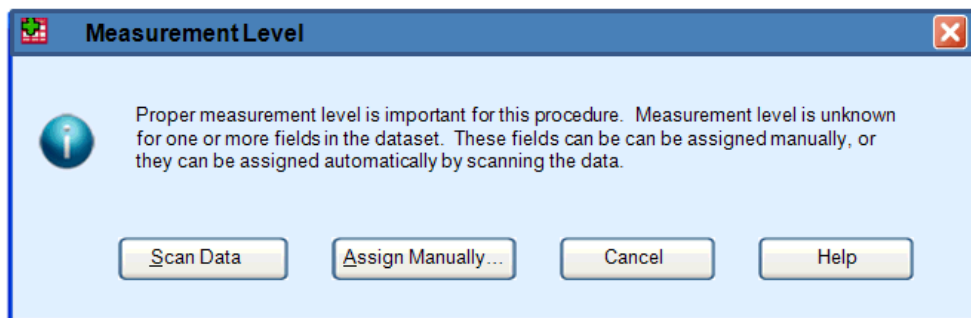
- ■ **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

- ■ **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Identify Unusual Cases Output

Figure 5-3
Identify Unusual Cases dialog box, Output tab
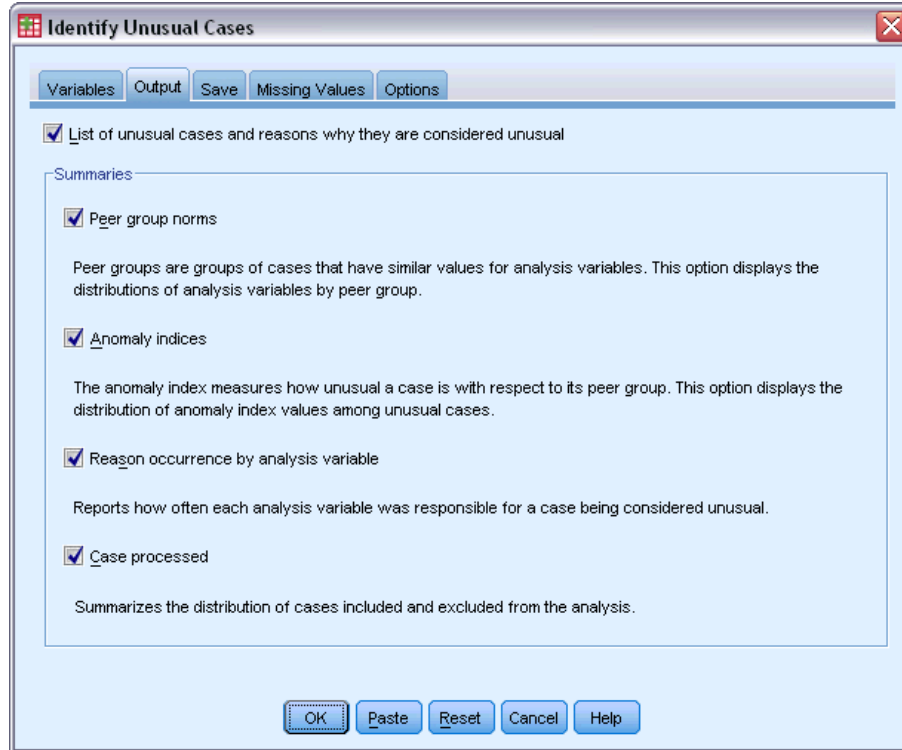
List of unusual cases and reasons why they are considered unusual. This option produces three tables:

■ The anomaly case index list displays cases that are identified as unusual and displays their corresponding anomaly index values.

■ The anomaly case peer ID list displays unusual cases and information concerning their corresponding peer groups.

■ The anomaly reason list displays the case number, the reason variable, the variable impact value, the value of the variable, and the norm of the variable for each reason.

All tables are sorted by anomaly index in descending order. Moreover, the IDs of the cases are displayed if the case identifier variable is specified on the Variables tab.

Summaries. The controls in this group produce distribution summaries.

■ **Peer group norms.** This option displays the continuous variable norms table (if any continuous variable is used in the analysis) and the categorical variable norms table (if any categorical variable is used in the analysis). The continuous variable norms table displays the mean and standard deviation of each continuous variable for each peer group. The categorical variable norms table displays the mode (most popular category), frequency, and frequency percentage of each categorical variable for each peer group. The mean of a continuous variable and the mode of a categorical variable are used as the norm values in the analysis.

■ **Anomaly indices.** The anomaly index summary displays descriptive statistics for the anomaly index of the cases that are identified as the most unusual.

■ **Reason occurrence by analysis variable.** For each reason, the table displays the frequency and frequency percentage of each variable's occurrence as a reason. The table also reports the descriptive statistics of the impact of each variable. If the maximum number of reasons is set to 0 on the Options tab, this option is not available.

■ **Cases processed.** The case processing summary displays the counts and count percentages for all cases in the active dataset, the cases included and excluded in the analysis, and the cases in each peer group.

Identify Unusual Cases Save

Figure 5-4
Identify Unusual Cases dialog box, Save tab

Save Variables. Controls in this group allow you to save model variables to the active dataset. You can also choose to replace existing variables whose names conflict with the variables to be saved.

■ **Anomaly index.** Saves the value of the anomaly index for each case to a variable with the specified name.

- **Peer groups.** Saves the peer group ID, case count, and size as a percentage for each case to variables with the specified rootname. For example, if the rootname *Peer* is specified, the variables *Peerid*, *PeerSize*, and *PeerPctSize* are generated. *Peerid* is the peer group ID of the case, *PeerSize* is the group's size, and *PeerPctSize* is the group's size as a percentage.

- **Reasons.** Saves sets of reasoning variables with the specified rootname. A set of reasoning variables consists of the name of the variable as the reason, its variable impact measure, its own value, and the norm value. The number of sets depends on the number of reasons requested on the Options tab. For example, if the rootname *Reason* is specified, the variables *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k*, and *ReasonNorm_k* are generated, where *k* is the *k*th reason. This option is not available if the number of reasons is set to 0.

Export Model File. Allows you to save the model in XML format.

Identify Unusual Cases Missing Values

Figure 5-5
Identify Unusual Cases dialog box, Missing Values tab
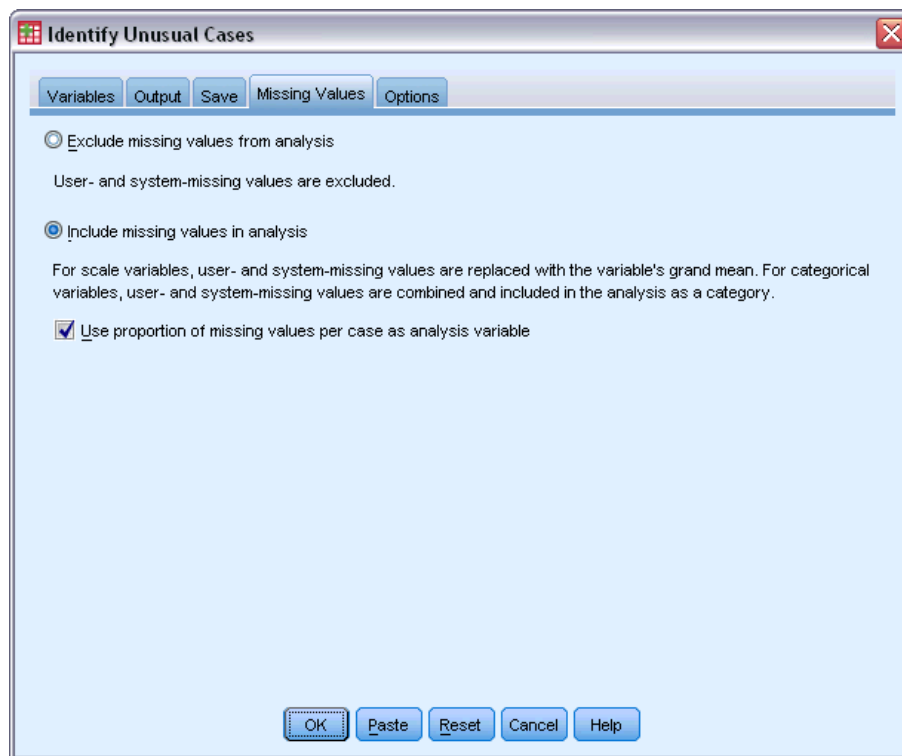
The Missing Values tab is used to control handling of user-missing and system-missing values.

- **Exclude missing values from analysis.** Cases with missing values are excluded from the analysis.

- **Include missing values in analysis.** Missing values of continuous variables are substituted with their corresponding grand means, and missing categories of categorical variables are grouped and treated as a valid category. The processed variables are then used in the analysis.

Optionally, you can request the creation of an additional variable that represents the proportion of missing variables in each case and use that variable in the analysis.

Identify Unusual Cases Options

Figure 5-6
Identify Unusual Cases dialog box, Options tab

Criteria for Identifying Unusual Cases. These selections determine how many cases are included in the anomaly list.

- **Percentage of cases with highest anomaly index values.** Specify a positive number that is less than or equal to 100.

- **Fixed number of cases with highest anomaly index values.** Specify a positive integer that is less than or equal to the total number of cases in the active dataset that are used in the analysis.

- **Identify only cases whose anomaly index value meets or exceeds a minimum value.** Specify a non-negative number. A case is considered anomalous if its anomaly index value is larger than or equal to the specified cutoff point. This option is used together with the Percentage of cases and Fixed number of cases options. For example, if you specify a fixed number of 50 cases and a cutoff value of 2, the anomaly list will consist of, at most, 50 cases, each with an anomaly index value that is larger than or equal to 2.

Number of Peer Groups. The procedure will search for the best number of peer groups between the specified minimum and maximum values. The values must be positive integers, and the minimum must not exceed the maximum. When the specified values are equal, the procedure assumes a fixed number of peer groups.

Note: Depending on the amount of variation in your data, there may be situations in which the number of peer groups that the data can support is less than the number specified as the minimum. In such a situation, the procedure may produce a smaller number of peer groups.

Maximum Number of Reasons. A reason consists of the variable impact measure, the variable name for this reason, the value of the variable, and the value of the corresponding peer group. Specify a non-negative integer; if this value equals or exceeds the number of processed variables that are used in the analysis, all variables are shown.

DETECTANOMALY Command Additional Features

The command syntax language also allows you to:

- Omit a few variables in the active dataset from analysis without explicitly specifying all of the analysis variables (using the EXCEPT subcommand).

- Specify an adjustment to balance the influence of continuous and categorical variables (using the MLWEIGHT keyword on the CRITERIA subcommand).

See the *Command Syntax Reference* for complete syntax information.

Optimal Binning

The Optimal Binning procedure discretizes one or more scale variables (referred to henceforth as **binning input variables**) by distributing the values of each variable into bins. Bin formation is optimal with respect to a categorical guide variable that "supervises" the binning process. Bins can then be used instead of the original data values for further analysis.

Examples. Reducing the number of distinct values a variable takes has a number of uses, including:

■ Data requirements of other procedures. Discretized variables can be treated as categorical for use in procedures that require categorical variables. For example, the Crosstabs procedure requires that all variables be categorical.

■ Data privacy. Reporting binned values instead of actual values can help safeguard the privacy of your data sources. The Optimal Binning procedure can guide the choice of bins.

■ Speed performance. Some procedures are more efficient when working with a reduced number of distinct values. For example, the speed of Multinomial Logistic Regression can be improved using discretized variables.

■ Uncovering complete or quasi-complete separation of data.

Optimal versus Visual Binning. The Visual Binning dialog boxes offer several automatic methods for creating bins without the use of a guide variable. These "unsupervised" rules are useful for producing descriptive statistics, such as frequency tables, but Optimal Binning is superior when your end goal is to produce a predictive model.

Output. The procedure produces tables of cutpoints for the bins and descriptive statistics for each binning input variable. Additionally, you can save new variables to the active dataset containing the binned values of the binning input variables and save the binning rules as command syntax for use in discretizing new data.

Data. This procedure expects the binning input variables to be scale, numeric variables. The guide variable should be categorical and can be string or numeric.

To Obtain Optimal Binning

From the menus choose:
Transform > Optimal Binning...

Figure 6-1
Optimal Binning dialog box, Variables tab
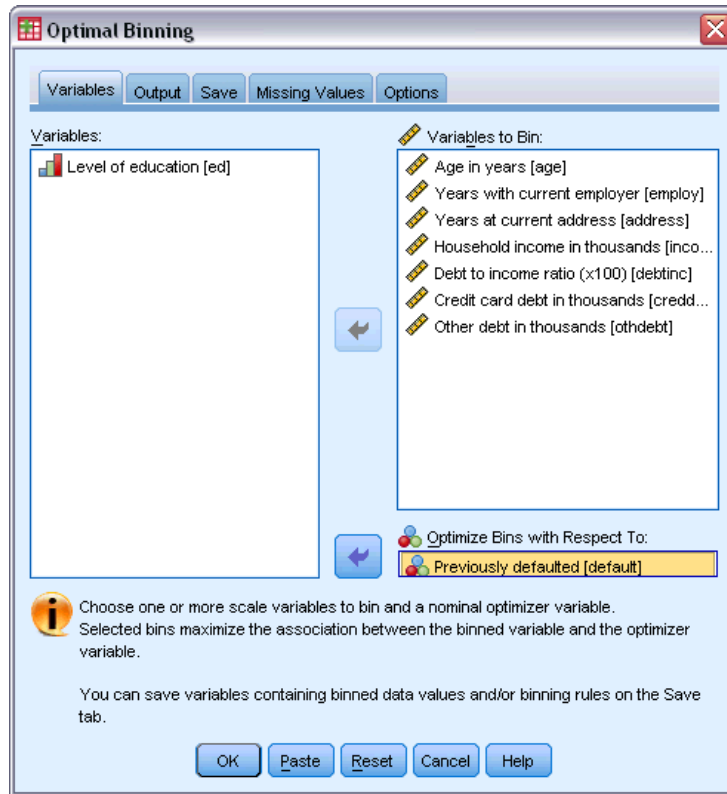
▶ Select one or more binning input variables.

▶ Select a guide variable.

Variables containing the binned data values are not generated by default. Use the Save tab to save these variables.

Optimal Binning Output

Figure 6-2
Optimal Binning dialog box, Output tab
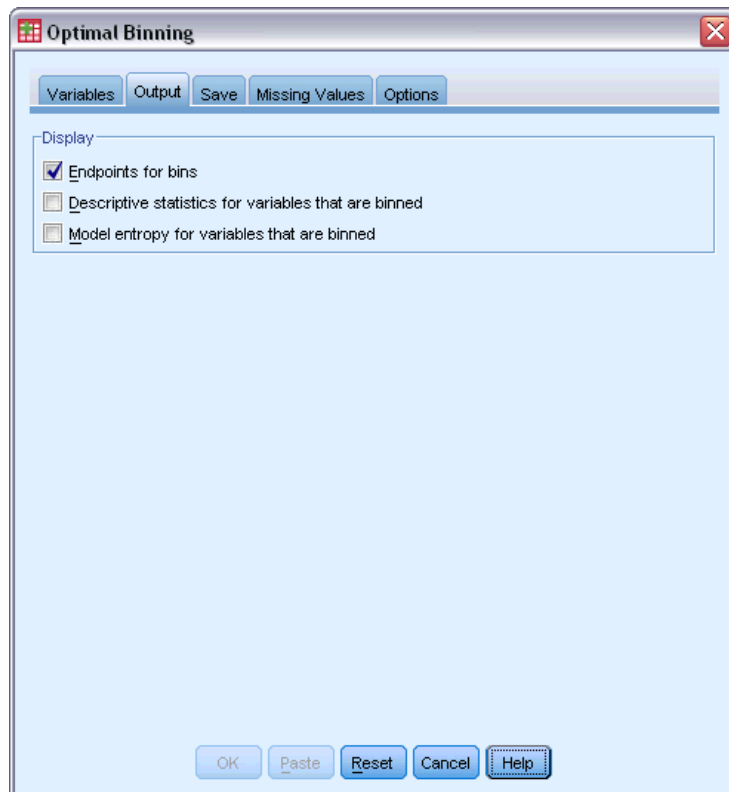
The Output tab controls the display of the results.

- **Endpoints for bins.** Displays the set of endpoints for each binning input variable.

- **Descriptive statistics for variables that are binned.** For each binning input variable, this option displays the number of cases with valid values, the number of cases with missing values, the number of distinct valid values, and the minimum and maximum values. For the guide variable, this option displays the class distribution for each related binning input variable.

- **Model entropy for variables that are binned.** For each binning input variable, this option displays a measure of the predictive accuracy of the variable with respect to the guide variable.

Optimal Binning Save

Figure 6-3
Optimal Binning dialog box, Save tab
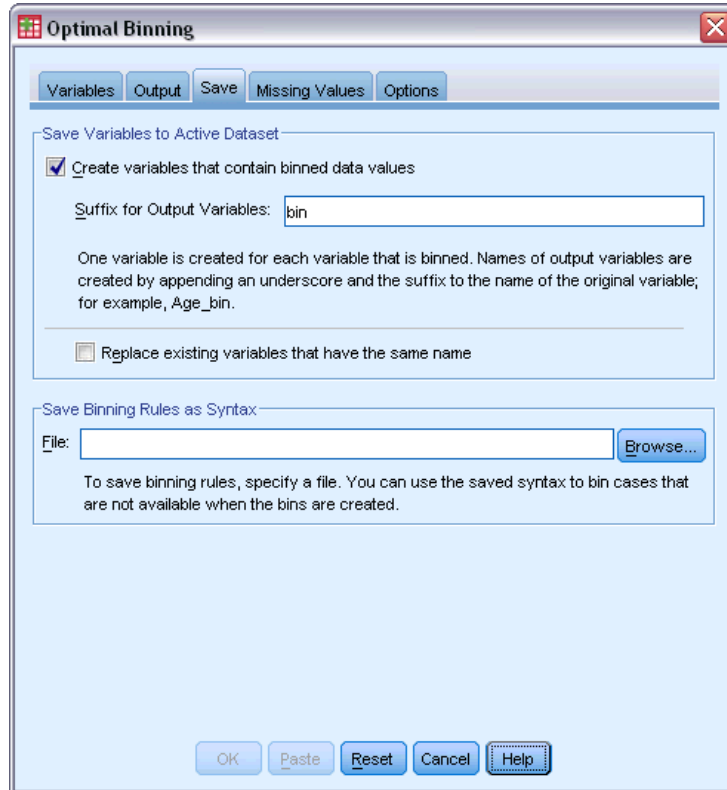
Save Variables to Active Dataset. Variables containing the binned data values can be used in place of the original variables in further analysis.

Save Binning Rules as Syntax. Generates command syntax that can be used to bin other datasets. The recoding rules are based on the cutpoints determined by the binning algorithm.

Optimal Binning Missing Values

Figure 6-4
Optimal Binning dialog box, Missing Values tab
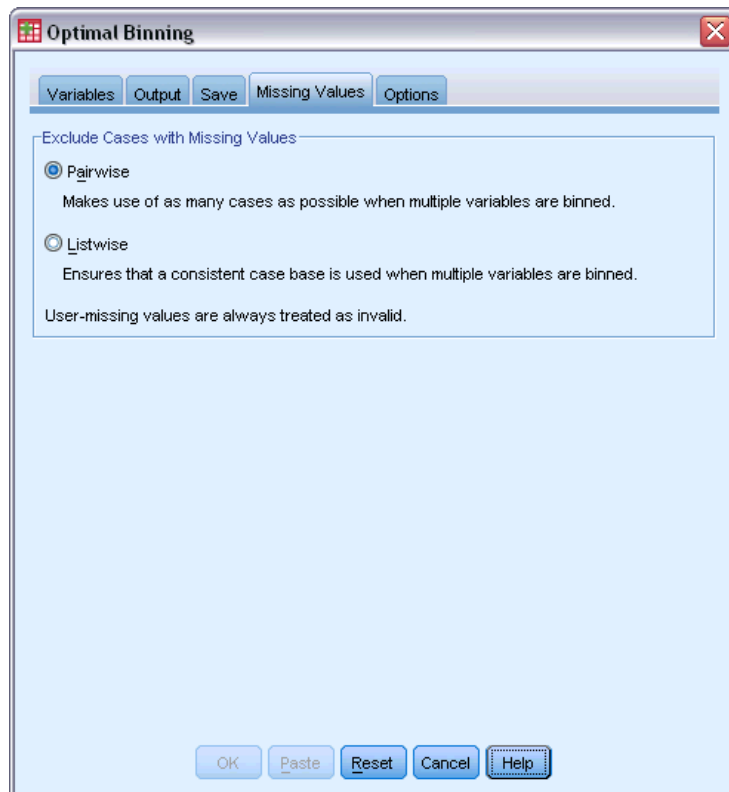
The Missing Values tab specifies whether missing values are handled using listwise or pairwise deletion. User-missing values are always treated as invalid. When recoding the original variable values into a new variable, user-missing values are converted to system-missing.

■ **Pairwise.** This option operates on each guide and binning input variable pair. The procedure will make use of all cases with nonmissing values on the guide and binning input variable.

■ **Listwise** This option operates across all variables specified on the Variables tab. If any variable is missing for a case, the entire case is excluded.

Optimal Binning Options

Figure 6-5
Optimal Binning dialog box, Options tab
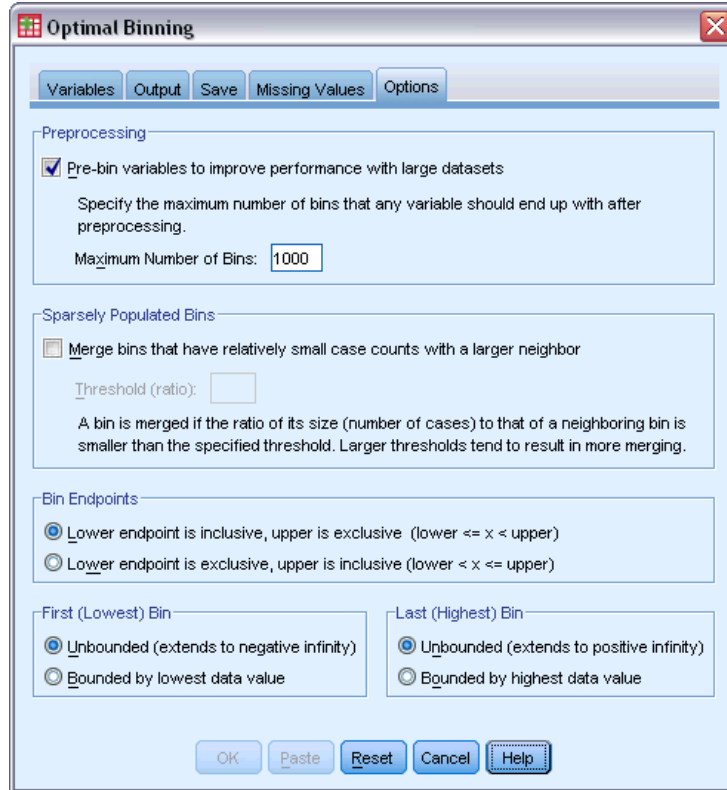
Preprocessing. "Pre-binning" binning input variables with many distinct values can improve processing time without a great sacrifice in the quality of the final bins. The maximum number of bins gives an upper bound on the number of bins created. Thus, if you specify 1000 as the maximum but a binning input variable has less than 1000 distinct values, the number of preprocessed bins created for the binning input variable will equal the number of distinct values in the binning input variable.

Sparsely Populated Bins. Occasionally, the procedure may produce bins with very few cases. The following strategy deletes these pseudo cutpoints:

▶ For a given variable, suppose that the algorithm found n_{final} cutpoints and thus $n_{final}+1$ bins. For bins $i = 2, ..., n_{final}$ (the second lowest-valued bin through the second highest-valued bin), compute

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

where *sizeof(b)* is the number of cases in the bin.

▶ When this value is less than the specified merging threshold, b_i is considered sparsely populated and is merged with b_{i-1} or b_{i+1}, whichever has the lower class information entropy.

The procedure makes a single pass through the bins.

Bin Endpoints. This option specifies how the lower limit of an interval is defined. Since the procedure automatically determines the values of the cutpoints, this is largely a matter of preference.

First (Lowest) / Last (Highest) Bin. These options specify how the minimum and maximum cutpoints for each binning input variable are defined. Generally, the procedure assumes that the binning input variables can take any value on the real number line, but if you have some theoretical or practical reason for limiting the range, you can bound it by the lowest / highest values.

OPTIMAL BINNING Command Additional Features

The command syntax language also allows you to:

- Perform unsupervised binning via the equal frequencies method (using the CRITERIA subcommand).

See the *Command Syntax Reference* for complete syntax information.

Part II: Examples

Validate Data

The Validate Data procedure identifies suspicious and invalid cases, variables, and data values.

Validating a Medical Database

An analyst hired by a medical group must maintain the quality of the information in the system. This process involves checking the values and variables and preparing a report for the manager of the data entry team.

The latest state of the database is collected in *stroke_invalid.sav*. For more information, see the topic Sample Files in Appendix A on p. 134. Use the Validate Data procedure to obtain the information that is necessary to produce the report. Syntax for producing these analyses can be found in *validatedata_stroke.sps*.

Performing Basic Checks

▶ To run a Validate Data analysis, from the menus choose:
Data > Validation > Validate Data...

Figure 7-1
Validate Data dialog box, Variables tab
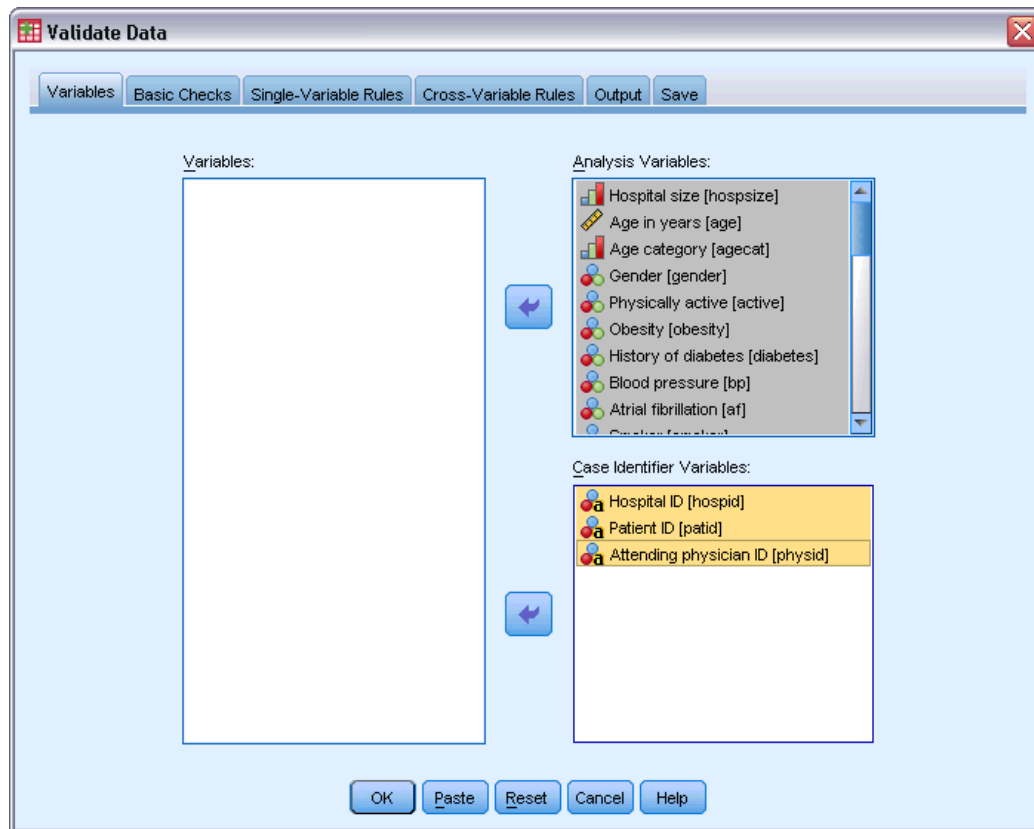
▶ Select *Hospital size* and *Age in years* through *Recoded Barthel index at 6 months* as analysis variables.

▶ Select *Hospital ID*, *Patient ID*, and *Attending physician ID* as case identifier variables.

▶ Click the Basic Checks tab.

Figure 7-2
Validate Data dialog box, Basic Checks tab

The default settings are the settings you want to run.

▶ Click OK.

Warnings

Figure 7-3
Warnings

The analysis variables passed the basic checks, and there are no empty cases, so a warning is displayed that explains why there is no output corresponding to these checks.

Incomplete Identifiers

Figure 7-4
Incomplete case identifiers

Case	Identifier		
	hospid	patid	physid
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

When there are missing values in case identification variables, the case cannot be properly identified. In this data file, case 288 is missing the *Patient ID*, while cases 573 and 774 are missing the *Hospital ID*.

Duplicate Identifiers

Figure 7-5
Duplicate case identifiers (first 11 shown)

Duplicate Identifiers Group	Number of Duplicates	Cases with Duplicate Identifiers	Identifier		
			hospid	patid	physid
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329

A case should be uniquely identified by the combination of values of the identifier variables. The first 11 entries in the duplicate identifiers table are shown here. These duplicates are patients with multiple events who were entered as separate cases for each event. Because this information can be collected in a single row, these cases should be cleaned up.

Copying and Using Rules from Another File

The analyst notes that the variables in this data file are similar to the variables from another project. The validation rules that are defined for that project are saved as properties of the associated data file and can be applied to this data file by copying the data properties of the file.

► To copy rules from another file, from the menus choose:
Data > Copy Data Properties...

Figure 7-6
Copy Data Properties, Step 1 (welcome)
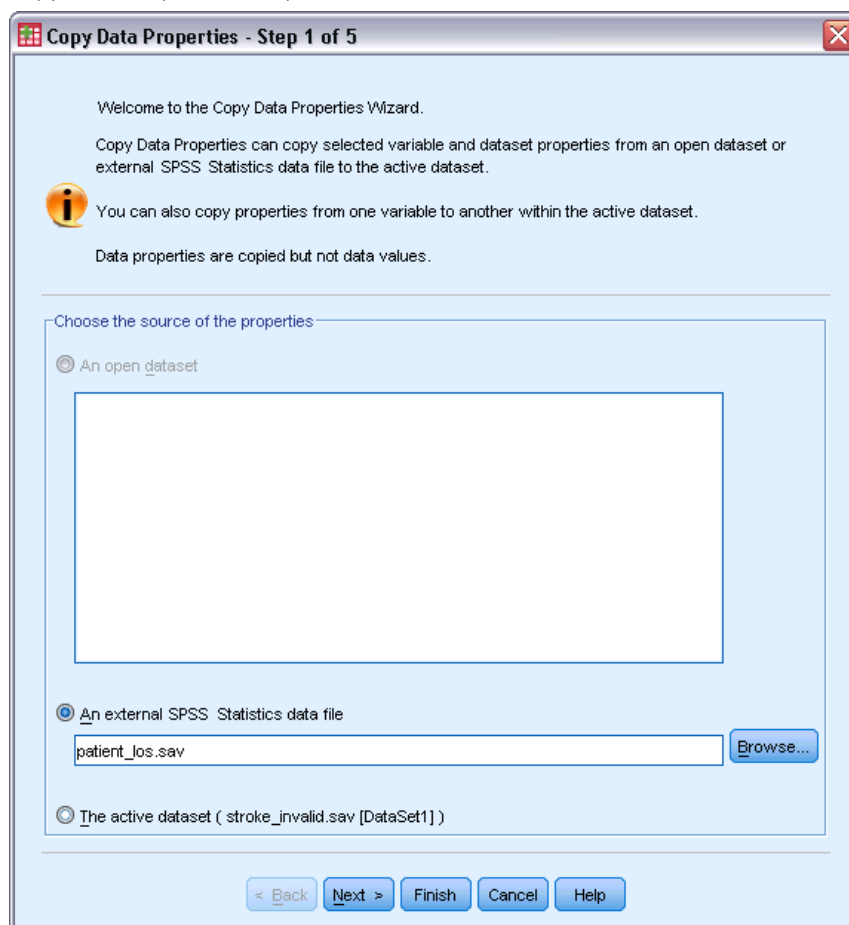
▶ Choose to copy properties from an external IBM® SPSS® Statistics data file, *patient_los.sav*. For more information, see the topic Sample Files in Appendix A on p. 134.

▶ Click Next.

Figure 7-7
Copy Data Properties, Step 2 (choose variables)
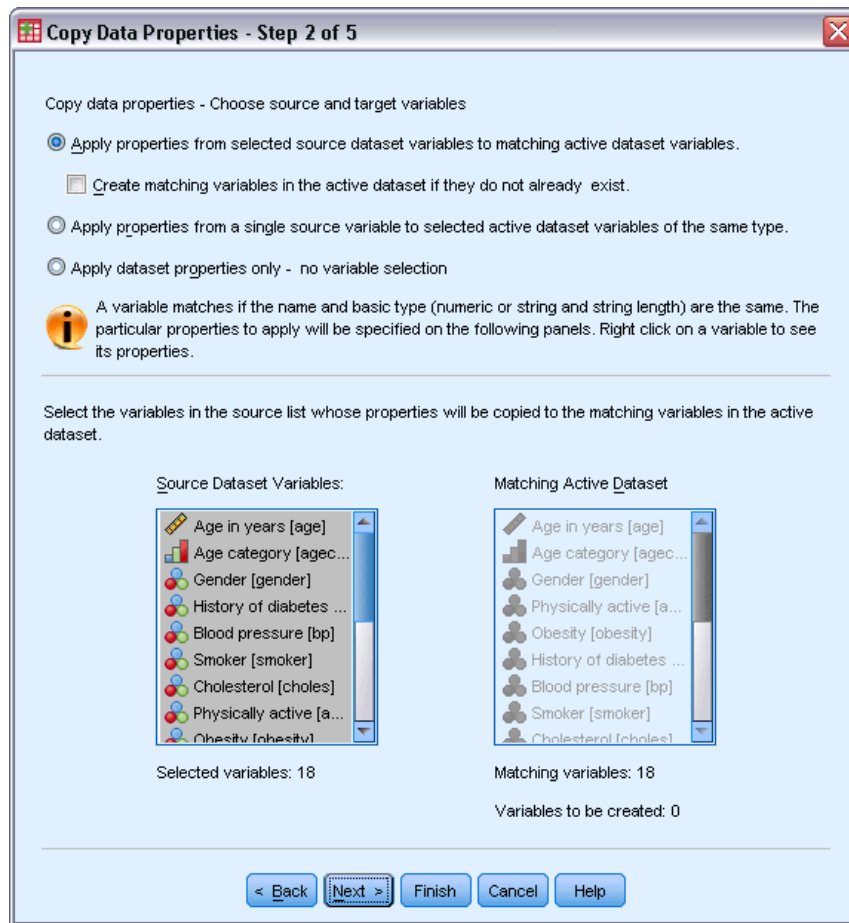
These are the variables whose properties you want to copy from *patient_los.sav* to the corresponding variables in *stroke_invalid.sav*.

▶ Click Next.

Figure 7-8
Copy Data Properties, Step 3 (choose variable properties)
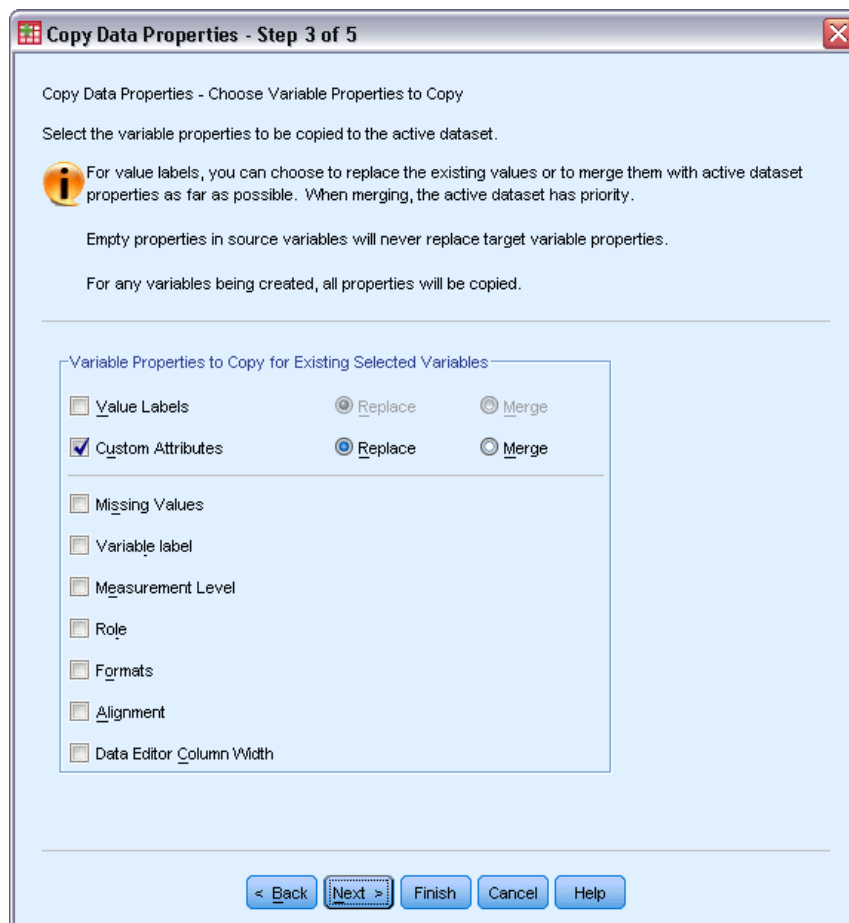
▶ Deselect all properties except Custom Attributes.

▶ Click Next.

Figure 7-9
Copy Data Properties, Step 4 (choose dataset properties)
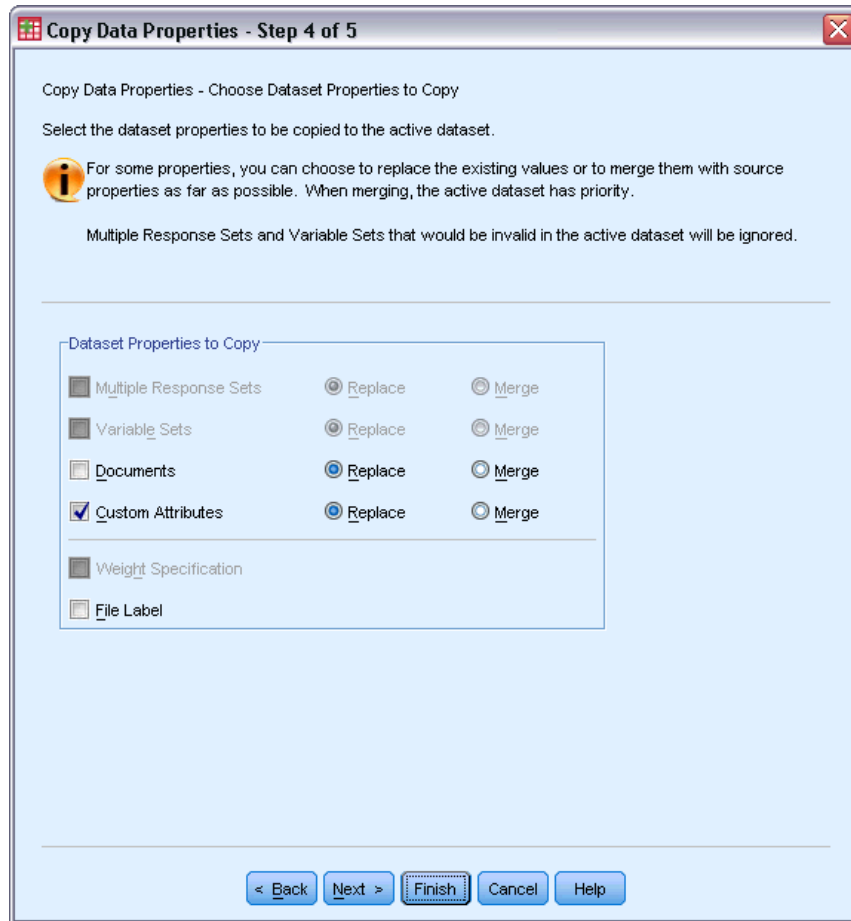
▶ Select Custom Attributes.

▶ Click Finish.

You are now ready to reuse the validation rules.

Figure 7-10
Validate Data dialog box, Single-Variable Rules tab
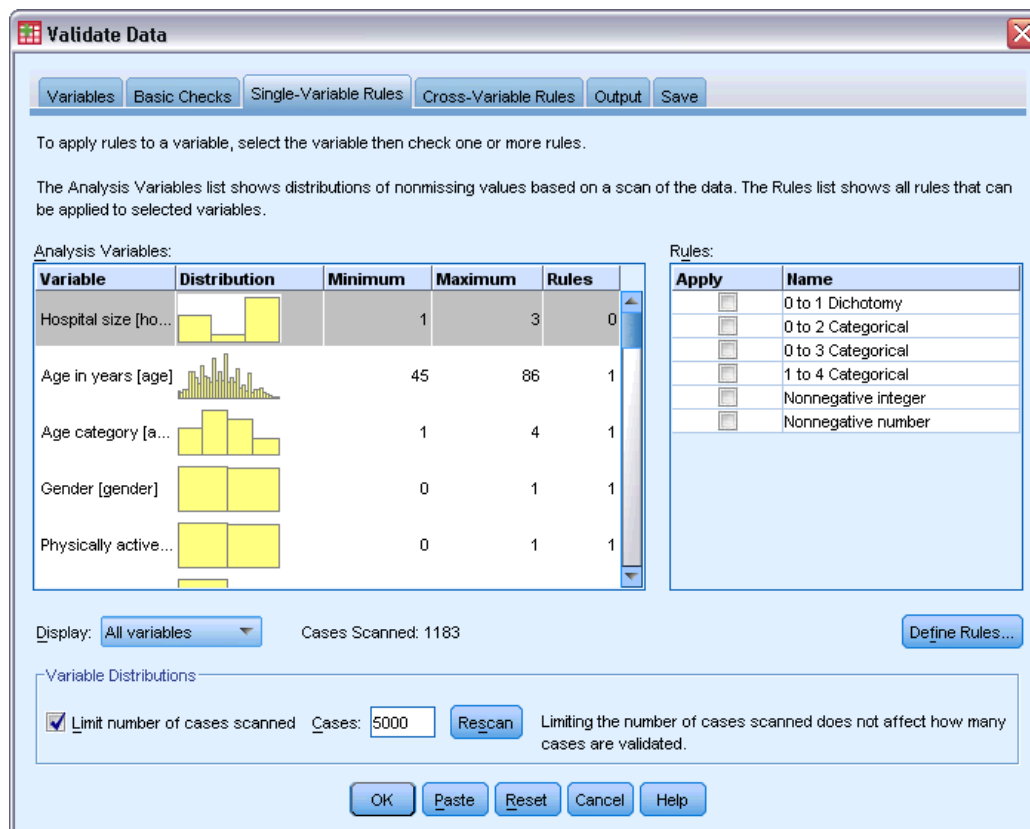
▶ To validate the *stroke_invalid.sav* data by using the copied rules, click the Dialog Recall toolbar button and choose Validate Data.

▶ Click the Single-Variable Rules tab.

The Analysis Variables list shows the variables that are selected on the Variables tab, some summary information about their distributions, and the number of rules attached to each variable. Variables whose properties were copied from *patient_los.sav* have rules that are attached to them.

The Rules list shows the single-variable validation rules that are available in the data file. These rules were all copied from *patient_los.sav*. Note that some of these rules are applicable to variables that did not have exact counterparts in the other data file.

Figure 7-11
Validate Data dialog box, Single-Variable Rules tab
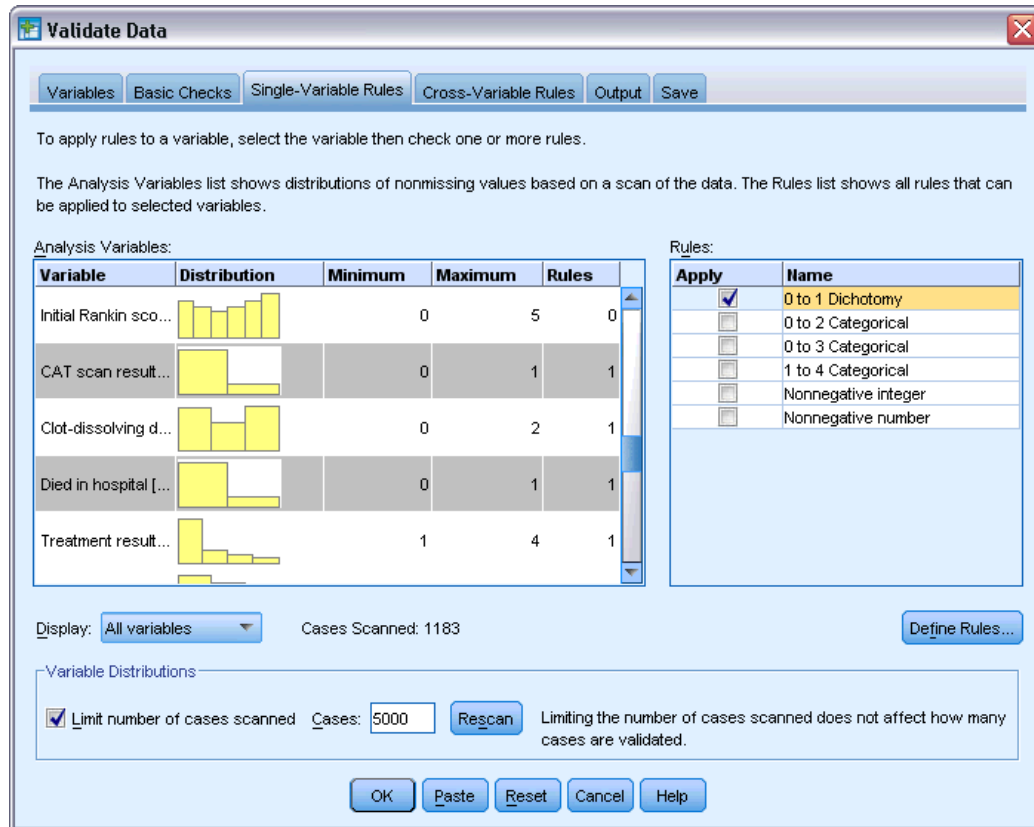
▶ Select *Atrial fibrillation*, *History of transient ischemic attack*, *CAT scan result*, and *Died in hospital* and apply the 0 to 1 Dichotomy rule.

▶ Apply 0 to 3 Categorical to *Post-event rehabilitation*.

▶ Apply 0 to 2 Categorical to *Post-event preventative surgery*.

▶ Apply Nonnegative integer to *Length of stay for rehabilitation*.

▶ Apply 1 to 4 Categorical to *Recoded Barthel index at 1 month* through *Recoded Barthel index at 6 months*.

▶ Click the Save tab.

Figure 7-12
Validate Data dialog box, Save tab
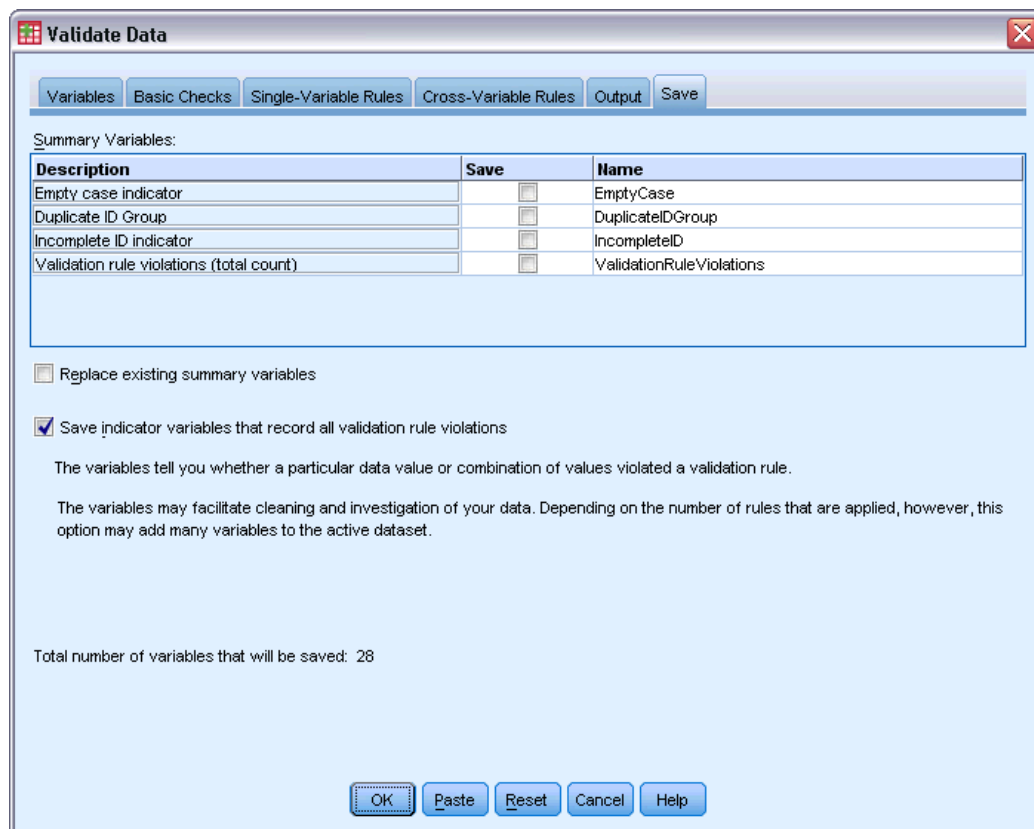
▶ Select Save indicator variables that record all validation rule violations. This process will make it easier to connect the case and variable that cause single-variable rule violations.

▶ Click OK.

Rule Descriptions

Figure 7-13
Rule descriptions

Rule	Description
Nonnegative integer	Type: Numeric Domain: Range Flag user-missing values: No Flag system-missing values: Yes Minimum: 0 Flag unlabeled values within range: No Flag noninteger values within range: Yes Rule: $VD.SRule[5]
0 to 1 Dichotomy	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 0, 1 Rule: $VD.SRule[1]
1 to 4 Categorical	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 1, 2, 3, 4 Rule: $VD.SRule[4]

Rules violated at least once are displayed.

The rule descriptions table displays explanations of rules that were violated. This feature is very useful for keeping track of a lot of validation rules.

Variable Summary

Figure 7-14
Variable summary

	Rule	Number of Violations
agecat	1 to 4 Categorical	1
	Total	1
gender	0 to 1 Dichotomy	1
	Total	1
angina	0 to 1 Dichotomy	1
	Total	1
time	Nonnegative integer	2
	Total	2
doa	0 to 1 Dichotomy	1
	Total	1

The variable summary table lists the variables that violated at least one validation rule, the rules that were violated, and the number of violations that occurred per rule and per variable.

Case Report

Figure 7-15
Case report

	Validation Rule	Identifier		
Case	Single-Variable[a]	hospid	patid	physid
175	0 to 1 Dichotomy (1)	OZN	0333204686	883285
274	0 to 1 Dichotomy (1)	OZN	1038840465	103254
310	Nonnegative integer (1)	OZN	2090290204	883285
437	0 to 1 Dichotomy (1)	WPA	2349729006	723384
752	Nonnegative integer (1)	GFG	4993307441	828754
1173	1 to 4 Categorical (1)	ALK	8737661990	185787

a. The number of variables that violated the rule follows each rule.

The case report table lists the cases (by both case number and case identifier) that violated at least one validation rule, the rules that were violated, and the number of times that the rule was violated by the case. The invalid values are shown in the Data Editor.

Figure 7-16
Data Editor with saved indicators of rule violations

	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity_	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_	@0to
1	4	.00	.00	.00	.00	.00	
2	4	.00	.00	.00	.00	.00	
3	1	.00	.00	.00	.00	.00	
4	4	.00	.00	.00	.00	.00	
5	3	.00	.00	.00	.00	.00	
6	4	.00	.00	.00	.00	.00	
7	4	.00	.00	.00	.00	.00	
8	4	.00	.00	.00	.00	.00	
9	4	.00	.00	.00	.00	.00	
10	2	.00	.00	.00	.00	.00	
11	2	.00	.00	.00	.00	.00	

Data View / Variable View /

A separate indicator variable is produced for each application of a validation rule. Thus, *@0to3Categorical_clotsolv_* is the application of the 0 to 3 Categorical single-variable validation rule to the variable *Clot-dissolving drugs*. For a given case, the easiest way to figure out which variable's value is invalid is simply to scan the values of the indicators. A value of 1 means that the associated variable's value is invalid.

Figure 7-17
Data Editor with indicator of rule violation for case 175

	recbart3	@0to1Dichot omy_doa_	@0to1Dichoto my_gender_	@0to1Dichoto my_angina	@1to4Categori cal_agecat_	Nonnegativeint eger_time_
172	4	.00	.00	.00	.00	.00
173	4	.00	.00	.00	.00	.00
174	3	.00	.00	.00	.00	.00
175	2	.00	.00	1.00	.00	.00
176	4	.00	.00	.00	.00	.00
177	3	.00	.00	.00	.00	.00
178	4	.00	.00	.00	.00	.00
179	3	.00	.00	.00	.00	.00
180	3	.00	.00	.00	.00	.00

Data View / Variable View /

Go to case 175, the first case with a rule violation. To speed your search, look at the indicators that are associated with variables in the variable summary table. It is easy to see that *History of angina* has the invalid value.

Figure 7-18
Data Editor with invalid value for History of angina

	af	smoker	choles	angina	mi	nitro	anticlot	tia
172	0	0	1	0	0	0	2	0
173	1	0	1	0	0	0	3	0
174	0	0	0	1	0	0	2	0
175	0	0	0	-1	1	0	1	0
176	0	0	0	0	0	0	0	0
177	0	0	0	0	0	0	0	0
178	0	0	1	0	0	0	0	0
179	0	0	0	0	0	0	1	0
180	0	0	0	0	0	0	0	1

Data View / Variable View /

History of angina has a value of −1. While this value is a valid missing value for treatment and result variables in the data file, it is invalid here because the patient history values do not currently have user-missing values defined.

Defining Your Own Rules

The validation rules that were copied from *patient_los.sav* have been very useful, but you need to define a few more rules to finish the job. Additionally, sometimes patients that are dead on arrival are accidentally marked as having died at the hospital. Single-variable validation rules cannot catch this situation, so you need to define a cross-variable rule to handle the situation.

▶ Click the Dialog Recall toolbar button and choose Validate Data.

▶ Click the Single-Variable Rules tab. (You need to define rules for *Hospital size*, the variables that measure Rankin scores, and the variables corresponding to the unrecoded Barthel indices.)

▶ Click Define Rules.

Figure 7-19
Define Validation Rules dialog box, Single-Variable Rules tab
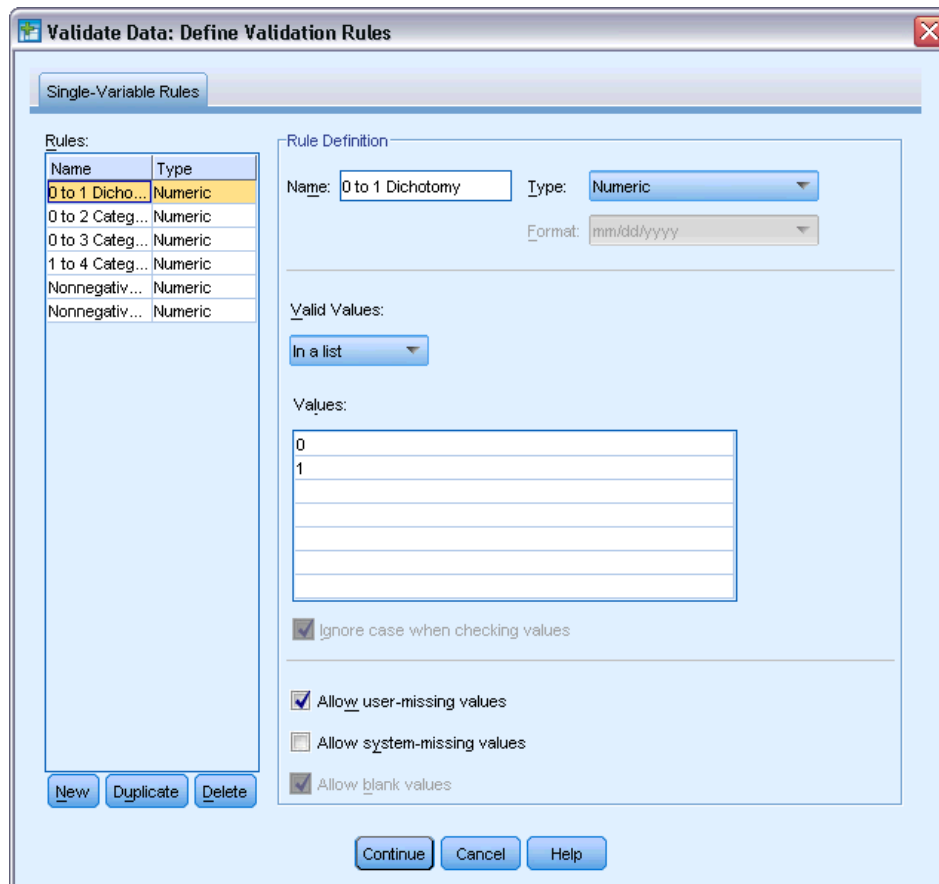
The currently defined rules are shown with 0 to 1 Dichotomy selected in the Rules list and the rule's properties displayed in the Rule Definition group.

▶ To define a rule, click New.

Figure 7-20
Define Validation Rules dialog box, Single-Variable Rules tab (1 to 3 Categorical defined)
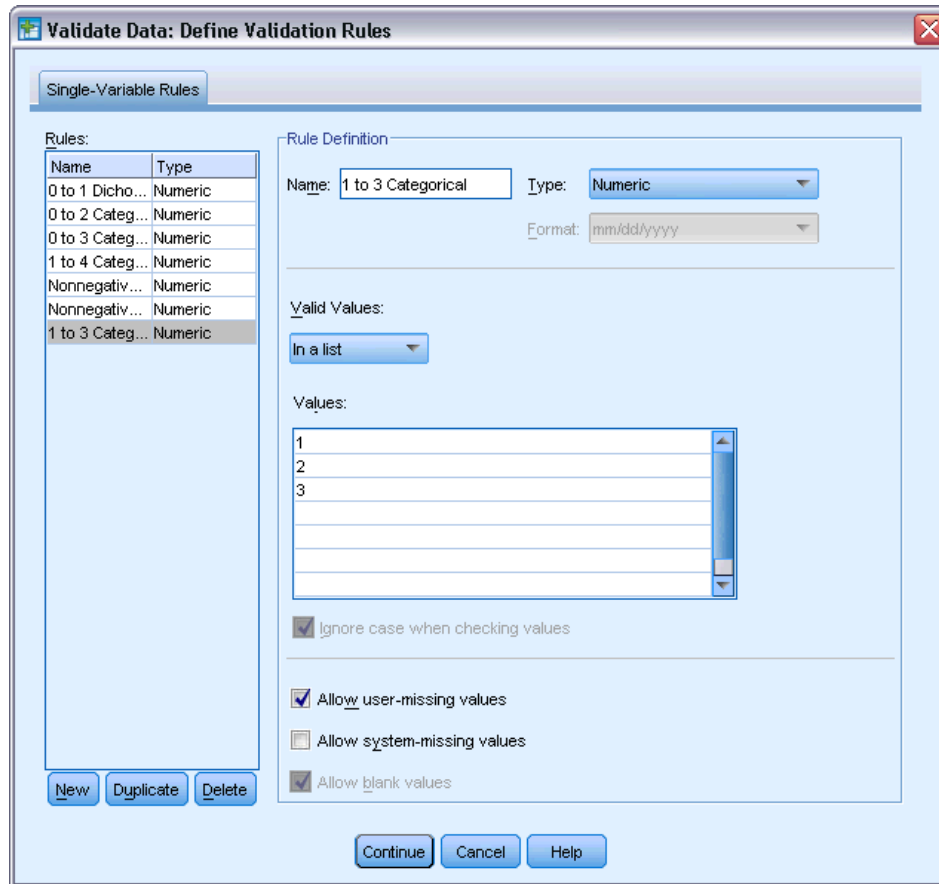
▶ Type 1 to 3 Categorical as the rule name.

▶ For Valid Values, choose In a list.

▶ Type 1, 2, and 3 as the values.

▶ Deselect Allow system-missing values.

▶ To define the rule for Rankin scores, click New.

Figure 7-21
Define Validation Rules dialog box, Single-Variable Rules tab (0 to 5 Categorical defined)
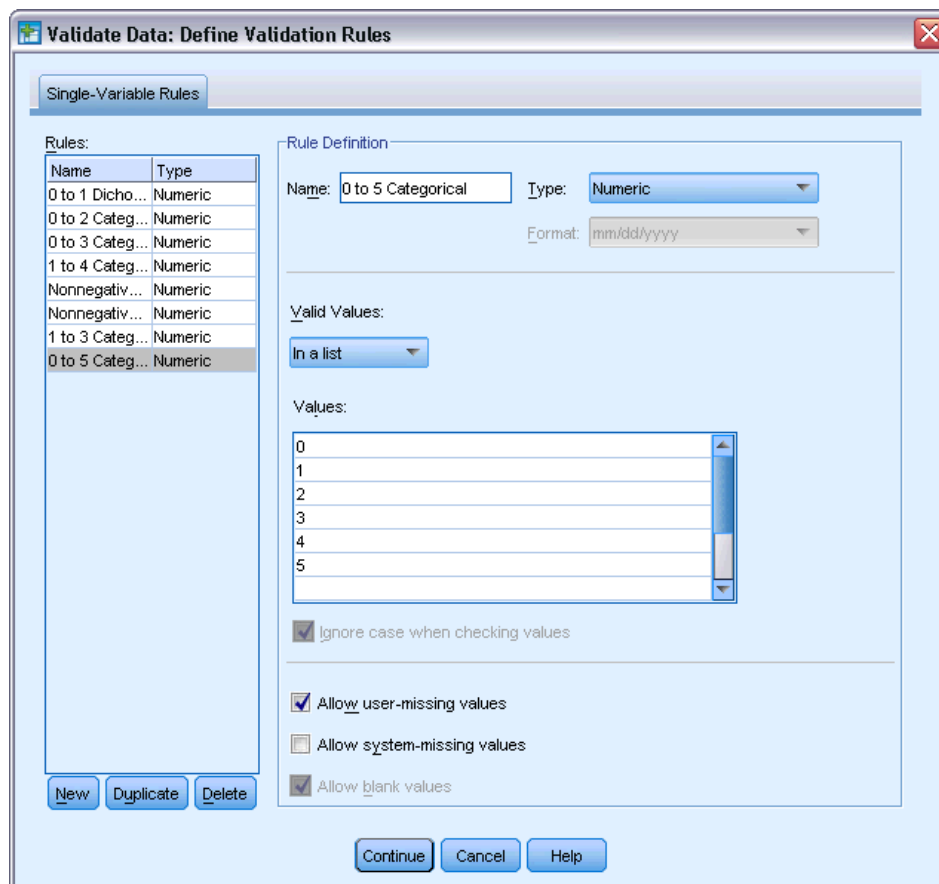
▶ Type 0 to 5 Categorical as the rule name.

▶ For Valid Values, choose In a list.

▶ Type 0, 1, 2, 3, 4, and 5 as the values.

▶ Deselect Allow system-missing values.

▶ To define the rule for Barthel indices, click New.

Figure 7-22
Define Validation Rules dialog box, Single-Variable Rules tab (0 to 100 by 5 defined)
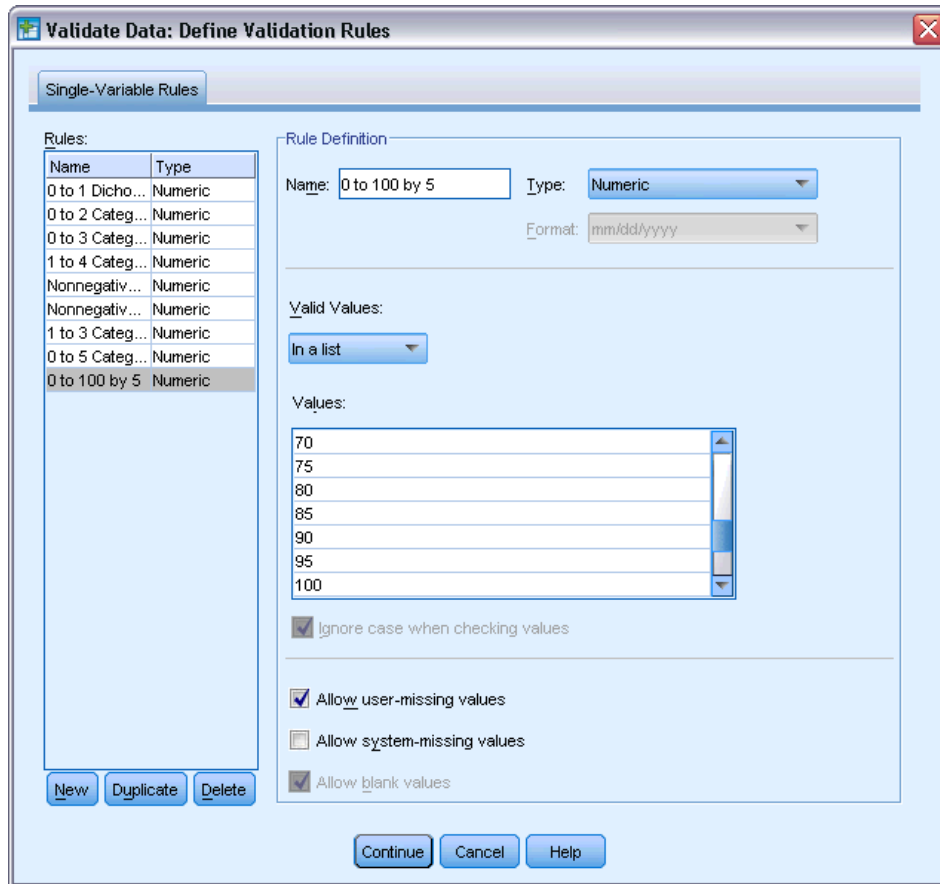
▶ Type 0 to 100 by 5 as the rule name.

▶ For Valid Values, choose In a list.

▶ Type 0, 5, ..., and 100 as the values.

▶ Deselect Allow system-missing values.

▶ Click Continue.

Figure 7-23
Validate Data dialog box, Single-Variable Rules tab (0 to 100 by 5 defined)
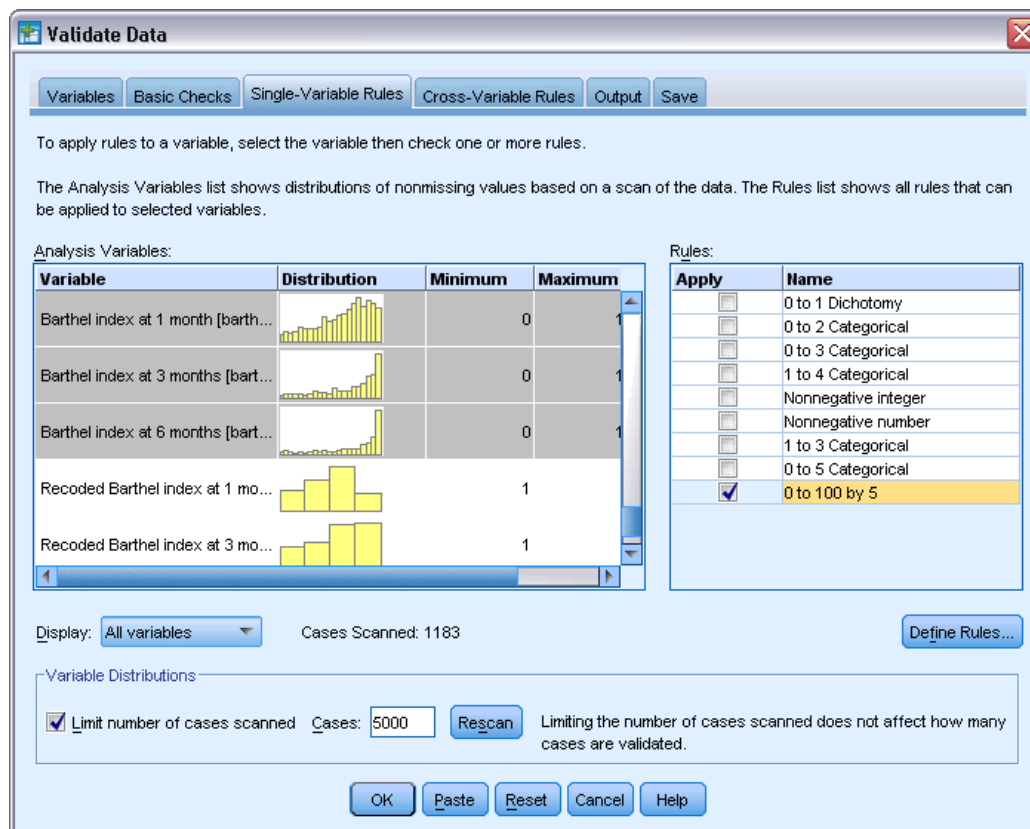
Now you need to apply the defined rules to analysis variables.

▶ Apply 1 to 3 Categorical to *Hospital size*.

▶ Apply 0 to 5 Categorical to *Initial Rankin score* and *Rankin score at 1 month* through *Rankin score at 6 months*.

▶ Apply 0 to 100 by 5 to *Barthel index at 1 month* through *Barthel index at 6 months*.

▶ Click the Cross-Variable Rules tab.

There are no currently defined rules.

▶ Click Define Rules.

Figure 7-24
Define Validation Rules dialog box, Cross-Variable Rules tab
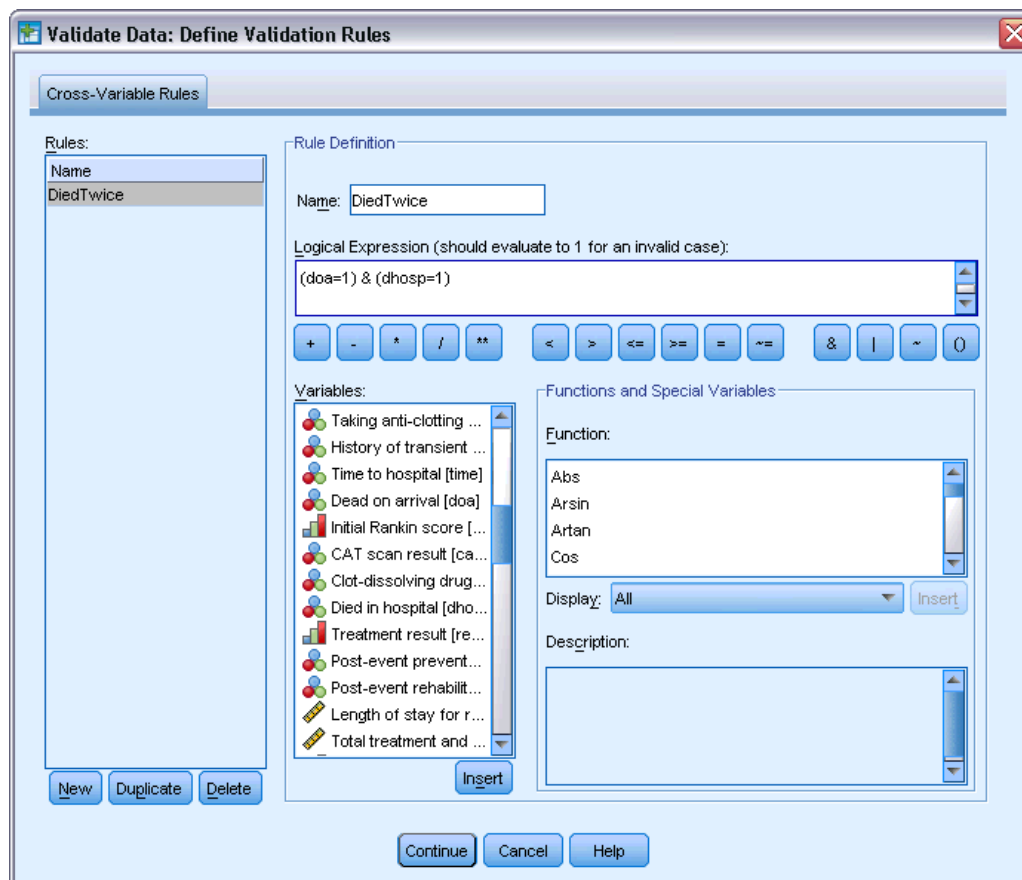
When there are no rules, a new placeholder rule is automatically created.

▶ Type **DiedTwice** as the name of the rule.

▶ Type **(doa=1) & (dhosp=1)** as the logical expression. This will return a value of 1 if the patient is recorded as both dead on arrival and died in the hospital.

▶ Click **Continue**.

The newly defined rule is automatically selected in the Cross-Variable Rules tab.

▶ Click **OK**.

Cross-Variable Rules

Figure 7-25
Cross-variable rules

Rule	Number of Violations	Rule Expression
DiedTwice	27	(doa=1) & (dhosp=1)

The cross-variable rules summary lists cross-variable rules that were violated at least once, the number of violations that occurred, and a description of each violated rule.

Case Report

Figure 7-26
Case report

Case	Validation Rule Violations		Identifier		
	Single-Variable[a]	Cross-Variable	hospid	patid	physid
20		Died twice	PBW	1192970826	355184
49		Died twice	NHV	8717862852	237418
129		Died twice	QWS	6901932085	215041
138		Died twice	RLD	1205005069	695521
162		Died twice	OZN	5546809538	125304
175	0 to 1 Dichotomy (1)		OZN	0333204686	883285
274	0 to 1 Dichotomy (1)		OZN	1038840465	103254
310	Nonnegative integer (1)		OZN	2090290204	883285
414		Died twice	WPA	3351107142	462020
437	0 to 1 Dichotomy (1)		WPA	2349729006	723384
447		Died twice	WPA	7163481282	519548
458		Died twice	WPA	9159094175	652070
462		Died twice	WPA	2137520354	723384
537		Died twice	SLB	5246122506	928076
544		Died twice	SLB	1605957462	506108
620		Died twice	GFG	8141858966	828754
629		Died twice	GFG	3397891610	539412
630		Died twice	GFG	3397891610	539412
639		Died twice	GFG	3962622031	327422
644		Died twice	GFG	4271782383	749432
649		Died twice	GFG	0950686750	618069
653		Died twice	GFG	0663642766	001448
722		Died twice	GFG	0418125590	877354
748		Died twice	GFG	8744721380	539412
752	Nonnegative integer (1) 0 to 1 Dichotomy (3)		GFG	4993307441	828754
868		Died twice	WWL	9714672452	237547
881		Died twice	WWL	6613279456	574275
915		Died twice	EFX	2575793702	501318
933		Died twice	IZO	2807437472	680253
1010		Died twice	BLA	5284009939	657638
1028		Died twice	BLA	8021997463	185703
1054		Died twice	ALK	0950897644	267830
1173	1 to 4 Categorical (1)		ALK	8737661990	185787

a. The number of variables that violated the rule follows each rule.

The case report now includes the cases that violated the cross-variable rule, as well as the previously discovered cases that violated single-variable rules. These cases all need to be reported to data entry for correction.

Summary

The analyst has the necessary information for a preliminary report to the data entry manager.

Related Procedures

The Validate Data procedure is a useful tool for data quality control.

■ The Identify Unusual Cases procedure analyzes patterns in your data and identifies cases with a few significant values that vary from type.

Automated Data Preparation

Preparing data for analysis is one of the most important steps in any project—and traditionally, one of the most time consuming. Automated Data Preparation (ADP) handles the task for you, analyzing your data and identifying fixes, screening out fields that are problematic or not likely to be useful, deriving new attributes when appropriate, and improving performance through intelligent screening techniques. You can use the algorithm in fully **automatic** fashion, allowing it to choose and apply fixes, or you can use it in **interactive** fashion, previewing the changes before they are made and accept or reject them as desired.

Using ADP enables you to make your data ready for model building quickly and easily, without needing prior knowledge of the statistical concepts involved. Models will tend to build and score more quickly; in addition, using ADP improves the robustness of automated modeling processes.

Using Automated Data Preparation Interactively

An insurance company with limited resources to investigate homeowners' insurance claims wants to build a model for flagging suspicious, potentially fraudulent claims. They have a sample of information on previous claims collected in *insurance_claims.sav*. For more information, see the topic Sample Files in Appendix A on p. 134. Before building the model, they will ready the data for modeling using automated data preparation. Since they want to be able to review the proposed transformations before the transformations are applied, they will use automated data preparation in interactive mode.

Choosing Between Objectives

▶ To run Automated Data Preparation interactively, from the menus choose:
Transform > Prepare Data for Modeling > Interactive...

Figure 8-1
Objective tab
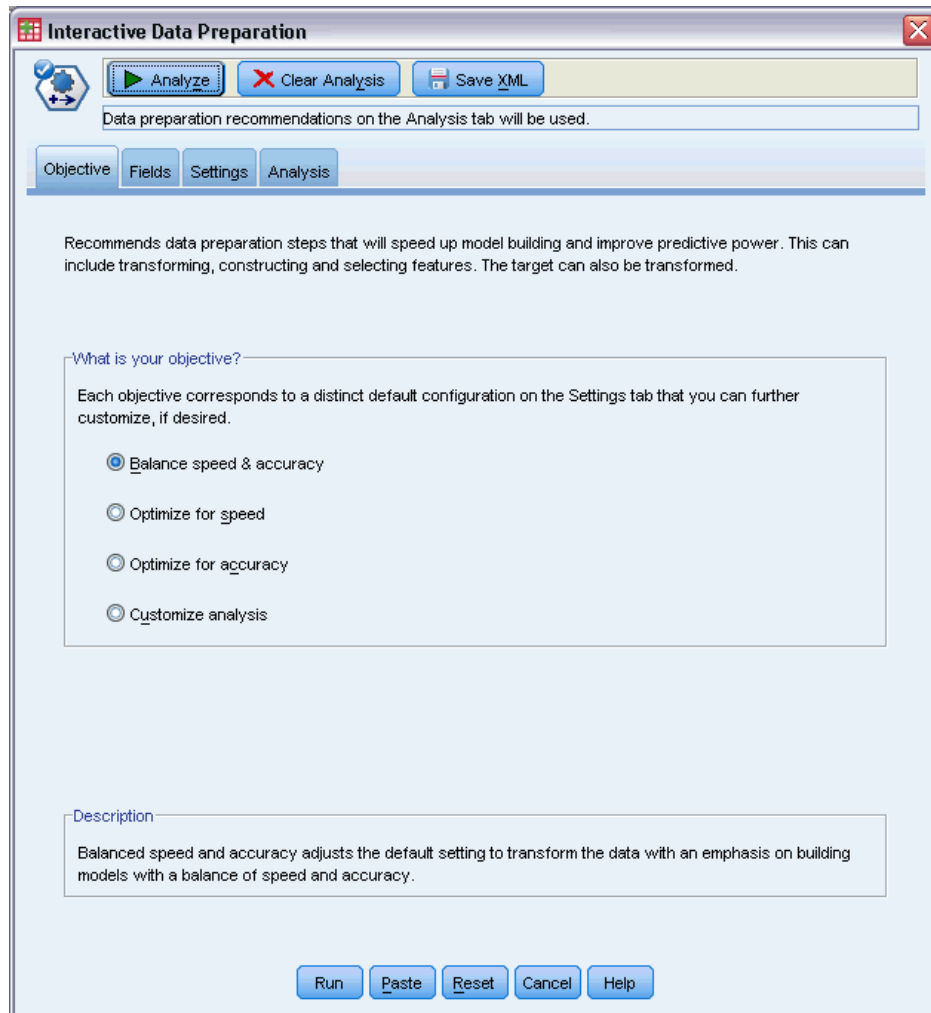
The first tab asks for an objective that controls the default settings, but what is the practical difference between the objectives? By running the procedure using each of the objectives we can see how the results differ.

▶ Make sure Balance speed & accuracy is selected and click Analyze.

Figure 8-2
Analysis tab, field processing summary for balanced objective

Field Processing Summary

Fields		N
Target		1
Input features		18
Features recommended for use in analysis	**Total**	18
	Original fields (untransformed)	9
	Transformations of original fields	4
	Derived from dates and times	5
	Constructed	0
Input features not used		0

Focus automatically switches to the Analysis tab while the procedure processes the data. The default main view is of the Field Processing Summary, which gives you an overview of how the fields were processed by automated data preparation. There is a single target, 18 inputs, and 18 fields recommended for model building. Of the fields recommended for modeling, 9 are original input fields, 4 are transformations of original input fields, and 5 are derived from date and time fields.

Figure 8-3
Analysis tab, predictive power for balanced objective

Features Recommended for Use in Analysis

Predictive Power

Target: fraudulent_transformed
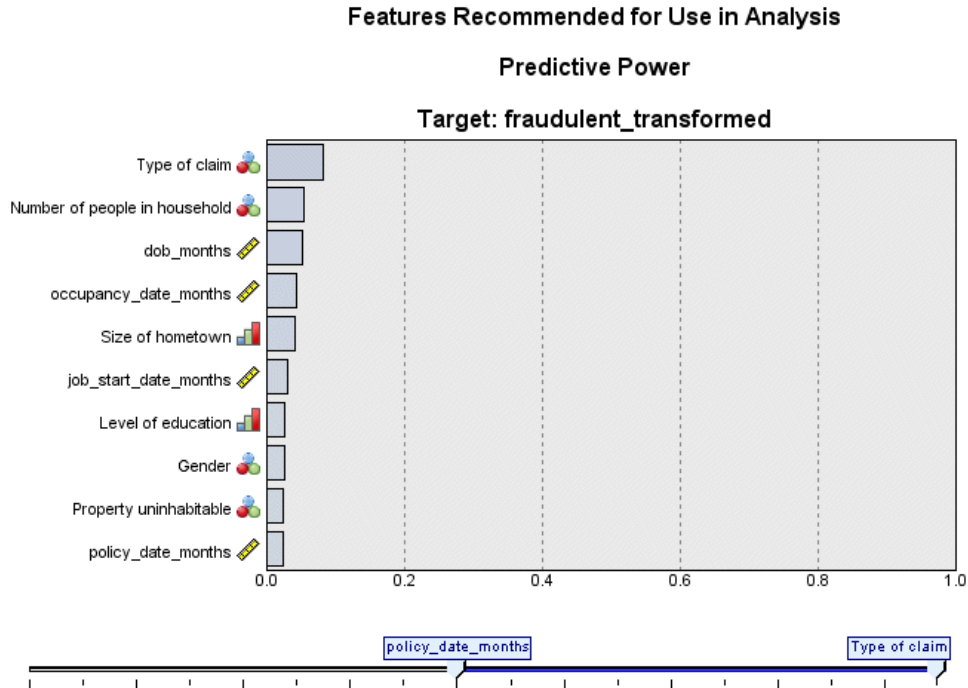
The default auxiliary view is of the Predictive Power, which quickly gives you an idea of which recommended fields will be most useful for model building. Note that while 18 predictors are recommended for analysis, only the first 10 are shown by default in the predictive power chart. To show more or fewer fields, use the slide control below the chart.

With Balance speed & accuracy as the objective, *Type of claim* is identified as the "best" predictor, followed by *Number of people in household* and the claimant's current age in months (the computed duration from the date of birth to the current date).

▶ Click Clear Analysis, then click on the Objective tab.

▶ Select Optimize for speed and click Analyze.

Figure 8-4
Analysis tab, field processing summary when optimized for speed

Field Processing Summary

Fields		N
Target		1
Input features		18
Features recommended for use in analysis	Total	2
	Original fields (untransformed)	0
	Transformations of original fields	2
	Derived from dates and times	0
	Constructed	0
Input features not used		16

- Feature construction was requested but no predictive features could be constructed. The most common reasons are: too few continuous input features that highly associated with the target or all continuous input features were independent.

Focus again automatically switches to the Analysis tab while the procedure processes the data. In this case, only 2 fields are recommended for model building, and both are transformations of the original fields.

Figure 8-5
Analysis tab, predictive power when optimized for speed
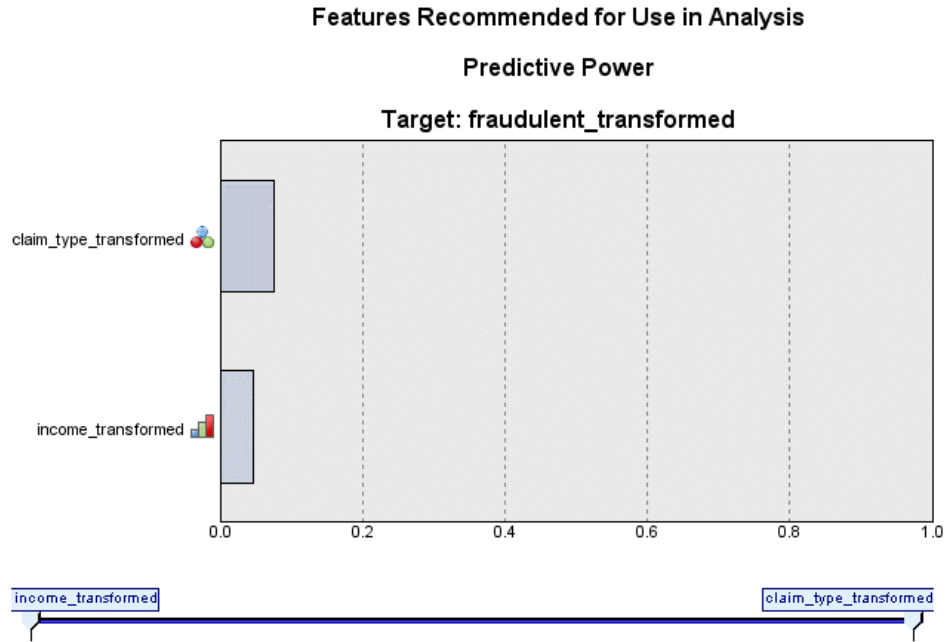
With Optimize for speed as the objective, *claim_type_transformed* is identified as the "best" predictor, followed by *income_transformed*.

▶ Click Clear Analysis, then click on the Objective tab.

▶ Select Optimize for accuracy and click Analyze.

Figure 8-6
Analysis tab, predictive power when optimized for accuracy

Field Processing Summary

Fields		N
Target		1
Input features		18
Features recommended for use in analysis	Total	32
	Original fields (untransformed)	9
	Transformations of original fields	4
	Derived from dates and times	19
	Constructed	0
Input features not used		0

With Optimize for accuracy as the objective, 32 fields are recommended for model building, as more fields are derived from dates and times by extracting days, months, and years from dates, and hours, minutes and seconds from times.

Figure 8-7
Analysis tab, predictive power when optimized for accuracy

Features Recommended for Use in Analysis

Predictive Power

Target: fraudulent_transformed
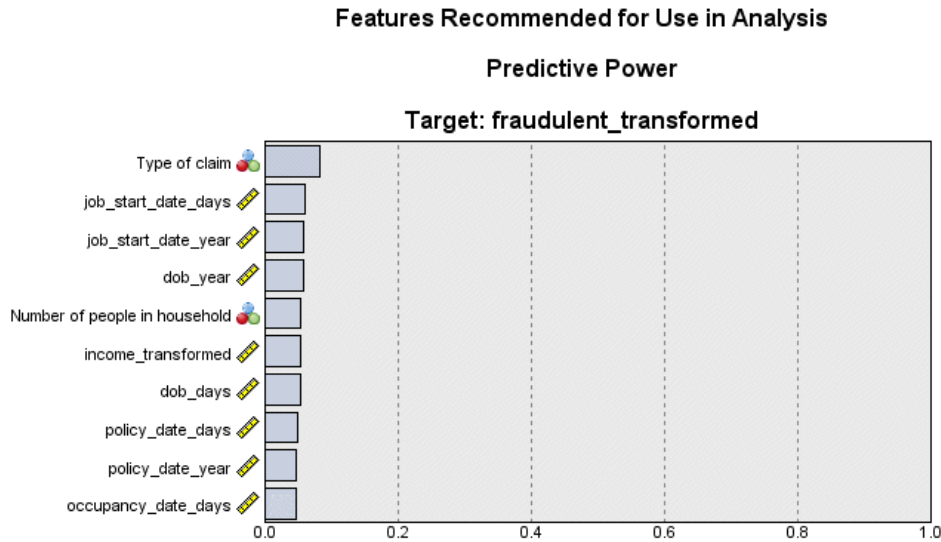
Type of claim is identified as the "best" predictor, followed by the number of days since the claimant started their most recent job (the computed duration from the job start date to the current date) and the year the claimant started their current job (extracted from the job start date).

To summarize:

■ Balance speed & accuracy creates fields usable in modeling from dates, and may transform continuous fields like *reside* to make them more normally distributed.

■ Optimize for accuracy creates some extra fields from dates (it also checks for outliers, and if the target is continuous, may transform it to make it more normally distributed).

■ Optimize for speed does not prepare dates and does not rescale continuous fields, but does merge categories of categorical predictors and bin continuous predictors when the target is categorical (and perform feature selection and construction when the target is continuous).

The insurance company decides to explore the Optimize for accuracy results further.

▶ Select Fields from the main view dropdown.

Fields and Field Details

Figure 8-8
Fields

Fields

Target

Name	Type
fraudulent	🔴

Features ☐ Include nonrecommended fields in table

Version to Use	Name	Type	Predictive Power
Original ▼	claim_type	🔴	0.08
Tran... ▼	job_start_date_days	📏	0.06
Tran... ▼	job_start_date_year	📏	0.06
Tran... ▼	dob_year	📏	0.06
Original ▼	reside	🔴	0.05
Tran... ▼	income	📏	0.05
Trans... ▼	dob_days	📏	0.05
Tran... ▼	policy_date_days	📏	0.05
Tran... ▼	policy_date_year	📏	0.05

The Fields view displays the processed fields and whether ADP recommends using them for model building. Clicking on any field name displays more information about the field in the linked view.

▶ Click income.

Figure 8-9
Field details for Household income in thousands

Details for Household income in thousands

The Field Details view shows the distributions of the original and transformed *Household income in thousands*. According to the processing table, records identified as outliers were trimmed (by setting their values equal to the cutoff for determining outliers) and the field was standardized to have mean 0 and standard deviation 1. The "bump" at the far right of the histogram for the transformed field shows that a number of records, perhaps more than 200, were identified as outliers. Income has a heavily skewed distribution, so this may be a case in which the default cutoff is too aggressive in determining outliers.

Also note the increase in predictive power of the transformed field over the original field. This appears to be a useful transformation.

► In the Fields view, click job_start_date_day. (Note that this is different from *job_start_date_days*.)

Figure 8-10
Field details for job_start_date_day

Details for job_start_date_day

Distribution

Skewness=-0.02
SD=1.00

Missing Values

0.0%

Predictive Power

Processing

Action	Details
Date and Time Features	Derive date elements: day
Check Type	Change type from ordered set to continuous
Outliers	Trim outliers
Continuous Features	Transform to standard units (Mean=0, SD=1)
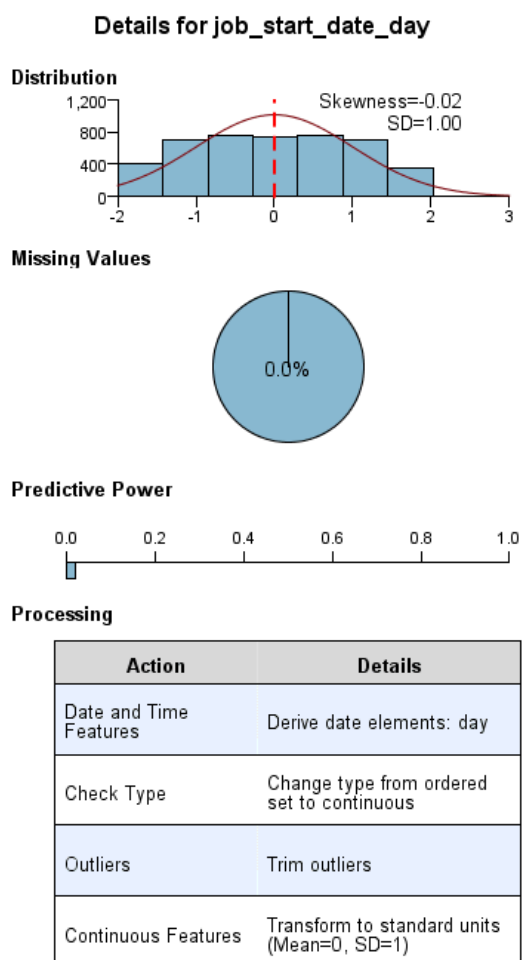
The field *job_start_date_day* is the day extracted from *Employment starting date [job_start_date]*. It is highly unlikely that this field has any real bearing on whether a claim is fraudulent, and so the insurance company wants to remove it from consideration for model building.

Figure 8-11
Field details for Household income in thousands

Trans... ▼	job_start_date_day	✎	0.02
Transformed Do not use	job_start_date_month	✎	0.02

▶ In the Fields view, select Do not use from the Version to Use dropdown in the job_start_date_day row. Perform the same operation for all fields with the *_day* and *_month* suffixes.

▶ To apply the transformations, click Run.

The dataset is now ready for model building, in the sense that all recommended predictors (both new and old) have their role set to Input, while non-recommended predictors have their role set to None. To create a dataset with only the recommended predictors, use the Apply Transformations settings in the dialog.

Using Automated Data Preparation Automatically

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over- and underperforming models, you want to establish a relationship between vehicle sales and vehicle characteristics. This information is collected in *car_sales_unprepared.sav*. For more information, see the topic Sample Files in Appendix A on p. 134. Use automated data preparation to prepare the data for analysis. Also build models using the data "before" and "after" preparation so that you can compare the results.

Preparing the Data

▶ To run automated data preparation in automatic mode, from the menus choose:
Transform > Prepare Data for Modeling > Automatic...

Figure 8-12
Objective tab
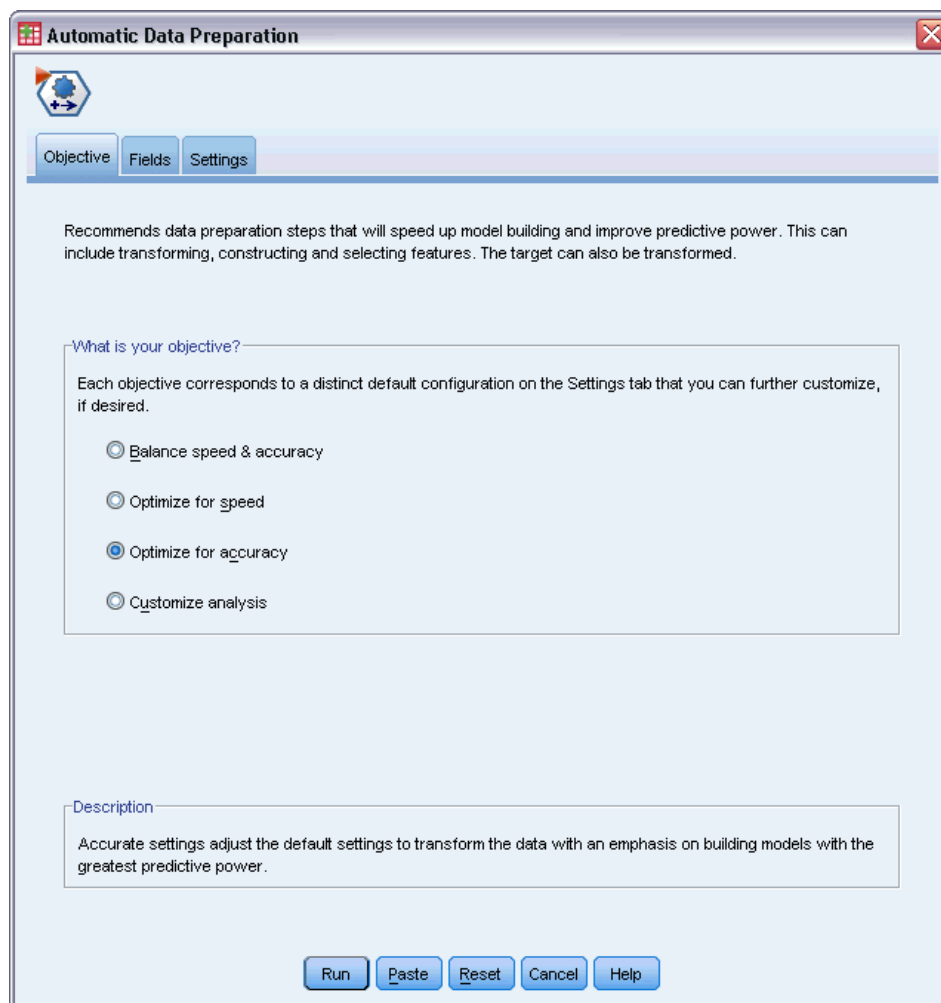
▶ Select Optimize for accuracy.

Since the target field, *Sales in thousands*, is continuous and could be transformed during automated data preparation, you want to save the transformations to an XML file in order to be able to use the Backtransform Scores dialog to convert predicted values of the transformed target back to the original scale.

▶ Click the Settings tab, then click on the Apply and Save settings.

Figure 8-13
Apply and Save settings
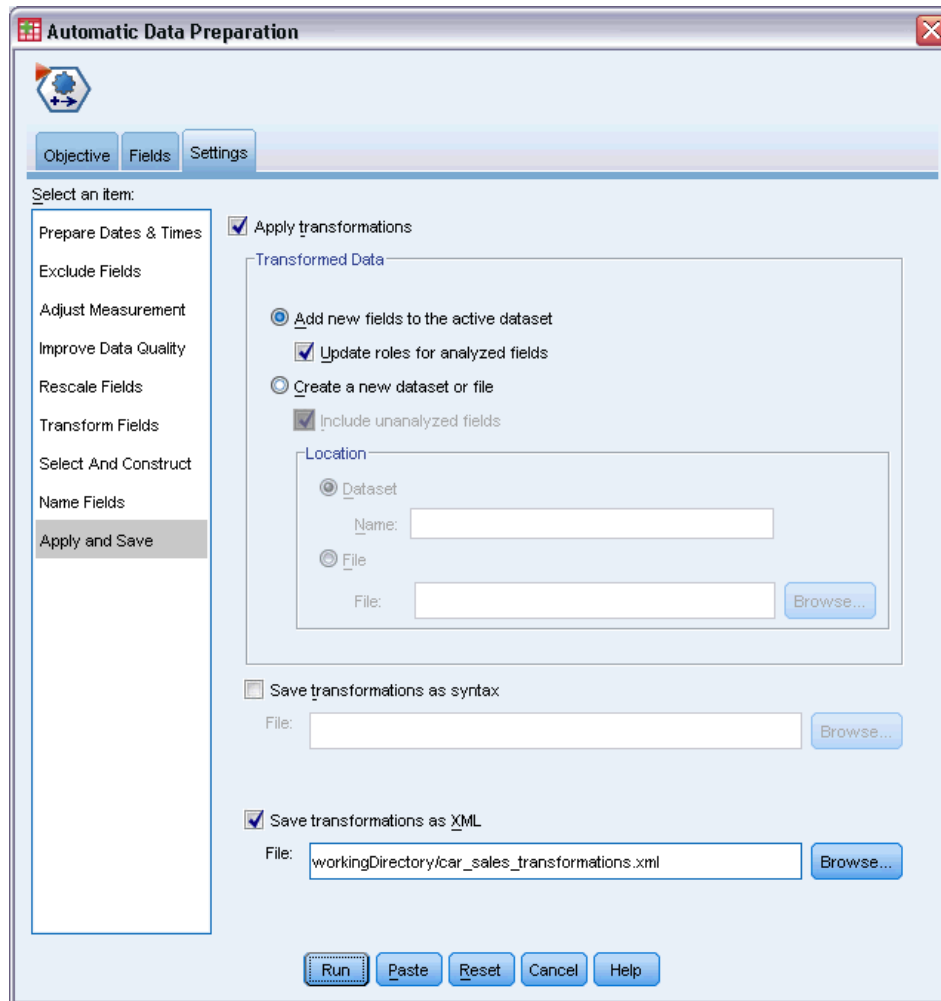
▶ Select Save transformations as XML and click Browse to navigate to
  workingDirectory/car_sales_transformations.xml, substituting the path you want to
  save the file to for workingDirectory.

▶ Click Run.

These selections generate the following command syntax:

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
```

```
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
  /REORDERNOMINAL INPUT=YES TARGET=NO
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
    CONSTRUCTION=NO
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

■ The `ADP` command prepares the target field *sales* and the input fields *resale* through *mpg*.

■ The `PREPDATETIME` subcommand is specified, but not used because none of the fields are date or time fields.

■ The `ADJUSTLEVEL` subcommand recasts ordinal fields with more than 10 values as continuous and continuous fields with fewer than 5 values as ordinal.

■ The `OUTLIERHANDLING` subcommand replaces values of continuous inputs (not the target) that are more than 3 standard deviations from the mean with the value that is 3 standard deviations from the mean.

■ The `REPLACEMISSING` subcommand replaces values of inputs (not the target) that are missing.

■ The `REORDERNOMINAL` subcommand recodes the values of nominal inputs from least frequently occurring to most frequently occurring.

■ The `RESCALE` subcommand standardizes continuous inputs to have mean 0 and standard deviation 1 using a z-score transformation, and standardizes the continuous target to have mean 0 and standard deviation 1 using a Box-Cox transformation.

■ The `TRANSFORM` subcommand turns off all default operations specified by this subcommand.

■ The `CRITERIA` subcommand specifies the default suffixes for transformations of the target and inputs.

■ The `OUTFILE` subcommand specifies that the transformations should be saved to */workingDirectory/car_sales_transformations.xml*, where */workingDirectory* is the path to where you want to save *car_sales_transformations.xml*.

■ The `TMS IMPORT` command reads the transformations in *car_sales_transformations.xml* and applies them to the active dataset, updating the roles of existing fields that are transformed.

■ The `EXECUTE` command causes the transformations to be processed. When using this as part of a longer stream of syntax, you may be able to remove the `EXECUTE` command to save some processing time.

Building a Model on the Unprepared Data

▶ To build a model on the unprepared data, from the menus choose:
Analyze > General Linear Model > Univariate...

Figure 8-14
GLM Univariate dialog
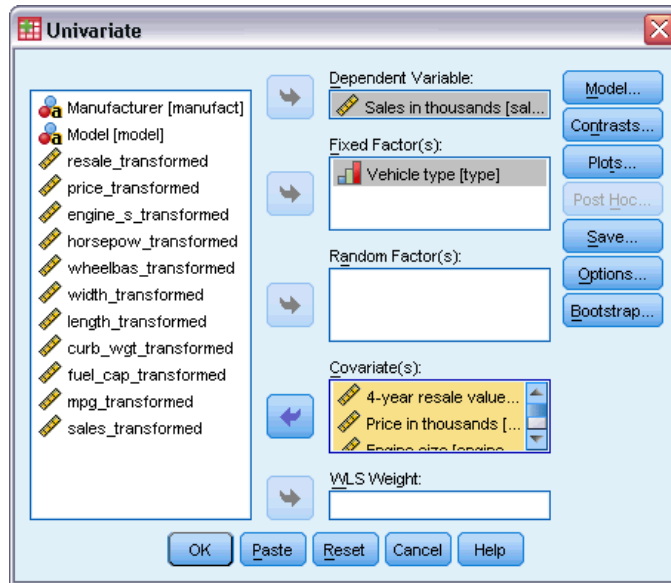
▶ Select *Sales in thousands [sales]* as the dependent variable.

▶ Select *Vehicle type [type]* as a fixed factor.

▶ Select *4-year resale value [resale]* through *Fuel efficiency [mpg]* as covariates.

▶ Click Save.

Figure 8-15
Save dialog
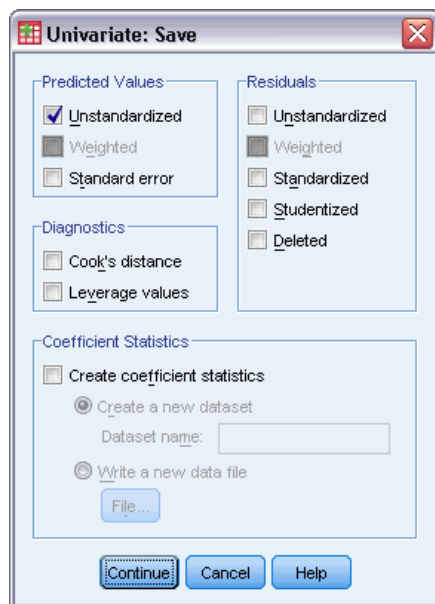
▶ Select Unstandardized in the Predicted Values group.

▶ Click Continue.

▶ Click OK in the GLM Univariate dialog box.

These selections generate the following command syntax:

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

Figure 8-16
Between-subjects effects for model based on unprepared data

Dependent Variable:Sales in thousands

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	226123.658[a]	11	20556.696	5.050	.000
Intercept	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
Error	427402.183	105	4070.497		
Total	1062354.955	117			
Corrected Total	653525.841	116			

a.R Squared = .346 (Adjusted R Squared = .277)

The default GLM Univariate output includes the between-subjects effects, which is an analysis of variance table. Each term in the model, plus the model as a whole, is tested for its ability to account for variation in the dependent variable. Note that variable labels are not displayed in this table.

The predictors show varying levels of significance; those with significance values less than 0.05 are typically considered useful to the model.

Building a Model on the Prepared Data

Figure 8-17
GLM Univariate dialog
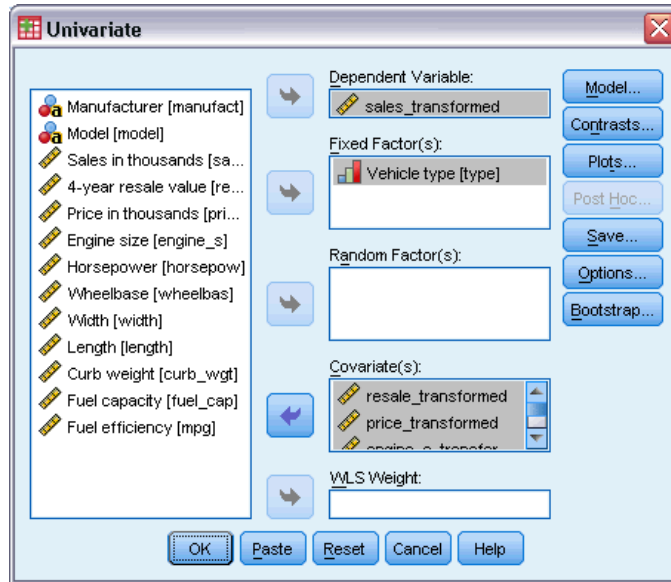
▶ To build the model on the prepared data, recall the GLM Univariate dialog.

▶ Deselect *Sales in thousands [sales]* and select *sales_transformed* as the dependent variable.

▶ Deselect *4-year resale value [resale]* through *Fuel efficiency [mpg]* and select *resale_transformed* through *mpg_transformed* as covariates.

▶ Click OK.

These selections generate the following command syntax:

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
    wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
    fuel_cap_transformed mpg_transformed type.
```

Figure 8-18
Between-subjects effects for model based on prepared data

Dependent Variable:sales_transformed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	78.965[a]	11	7.179	13.512	.000
Intercept	2.549	1	2.549	4.797	.030
resale_transformed	.852	1	.852	1.603	.207
price_transformed	8.540	1	8.540	16.075	.000
engine_s_transformed	2.943	1	2.943	5.540	.020
horsepow_transformed	.054	1	.054	.102	.749
wheelbas_transformed	1.148	1	1.148	2.161	.144
width_transformed	.026	1	.026	.049	.826
length_transformed	.407	1	.407	.766	.383
curb_wgt_transformed	.027	1	.027	.051	.822
fuel_cap_transformed	.089	1	.089	.168	.682
mpg_transformed	3.226	1	3.226	6.073	.015
type	4.268	1	4.268	8.033	.005
Error	77.035	145	.531		
Total	156.000	157			
Corrected Total	156.000	156			

a.R Squared = .506 (Adjusted R Squared = .469)

There are a few interesting differences to note in the between-subjects effects for the model built on the unprepared data and the model built on the prepared data. First, note that the total degrees of freedom has increased; this is due to the fact that missing values were replaced with imputed values during automated data preparation, so records that were listwise removed from the first model are available to the second. More notably, perhaps, the significance of certain predictors has changed. While both models agree that the engine size [*engine_s*] and vehicle type [*type*] are useful to the model, the wheelbase [*wheelbas*] and curb weight [*curb_wgt*] are no longer significant, and the vehicle price [*price_transformed*] and fuel efficiency [*mpg_transformed*] are now significant.

Why did this change occur? Sales has a skewed distribution, so it could be that wheelbase and curb weight had a few influential records that were no longer influential once sales was transformed. Another possibility is that the extra cases available due to missing value replacement changed the statistical significance of these variables. In any case, this would require further investigation that we will not pursue here.

Note that the R Squared is higher for the model built on the prepared data, but because sales has been transformed, this may not be the best measure for comparing each model's performance. Instead, you can compute the nonparametric correlations between the observed values and the two sets of predicted values.

Comparing the Predicted Values

▶ To obtain correlations of the predicted values from the two models, from the menus choose:
Analyze > Correlate > Bivariate...

Figure 8-19
Bivariate Correlations dialog
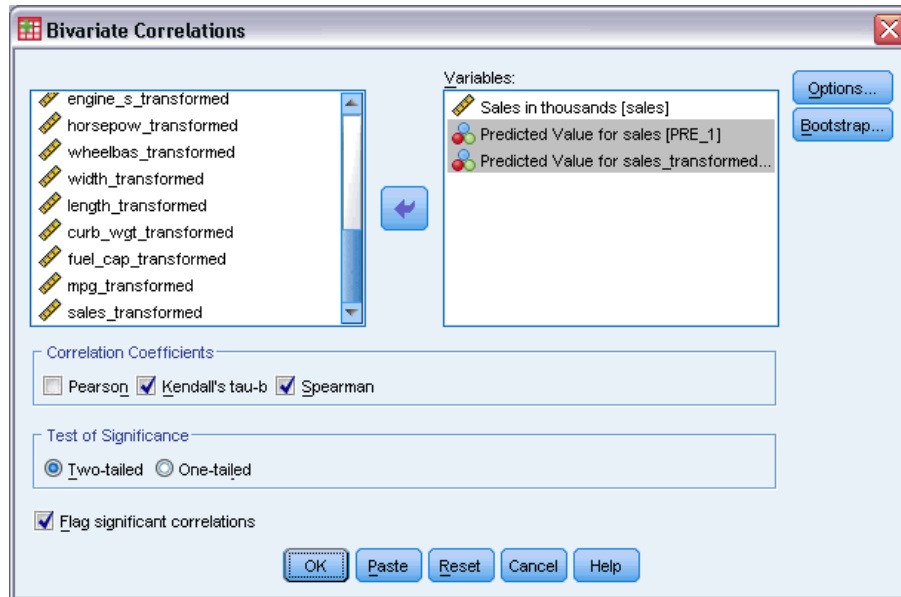
▶ Select *Sales in thousands [sales]*, *Predicted Value for sales [PRE_1]*, and *Predicted Values for sales_transformed [PRE_2]* as analysis variables.

▶ Deselect Pearson and select Kendall's tau-b and Spearman in the Correlation Coefficients group.

Note that the *Predicted Values for sales_transformed [PRE_2]* can be used to compute the nonparametric correlations without having to backtransform it to the original scale because the backtransformation does not change the rank order of the predicted values.

▶ Click OK.

These selections generate the following command syntax:

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Figure 8-20
Nonparametric correlations

			Sales in thousands	Predicted Value for sales	Predicted Value for sales_ transformed
Kendall's tau_b	Sales in thousands	Correlation Coefficient	1.000	.376**	.480**
		Sig. (2-tailed)	.	.000	.000
		N	157	117	157
	Predicted Value for sales	Correlation Coefficient	.376**	1.000	.659**
		Sig. (2-tailed)	.000	.	.000
		N	117	117	117
	Predicted Value for sales_transformed	Correlation Coefficient	.480**	.659**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	157	117	157
Spearman's rho	Sales in thousands	Correlation Coefficient	1.000	.530**	.664**
		Sig. (2-tailed)	.	.000	.000
		N	157	117	157
	Predicted Value for sales	Correlation Coefficient	.530**	1.000	.835**
		Sig. (2-tailed)	.000	.	.000
		N	117	117	117
	Predicted Value for sales_transformed	Correlation Coefficient	.664**	.835**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	157	117	157

**.Correlation is significant at the 0.01 level (2-tailed).

The first column shows that the predicted values for the model built using the prepared data are more strongly correlated with the observed values by both Kendall's tau-b and Spearman's rho measures. This suggests that running automated data preparation has improved the model.

Backtransforming the Predicted Values

▶ The prepared data includes a transformation of sales, so the predicted values from this model are not directly usable as scores. To transform the predicted values into the original scale, from the menus choose:

Transform > Prepare Data for Modeling > Backtransform Scores...

Figure 8-21
Backtransform Scores dialog
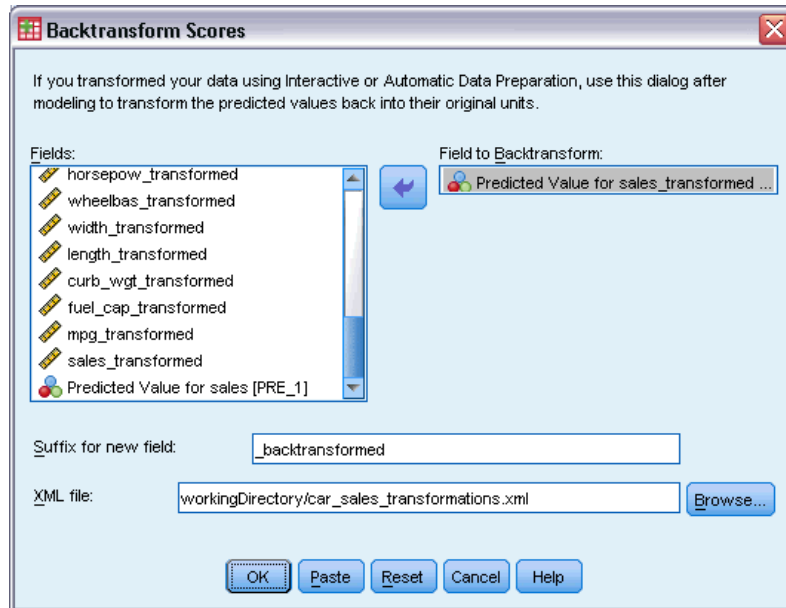
▶ Select *Predicted Value for sales_transformed [PRE_2]* as the field to backtransform.

▶ Type _backtransformed as the suffix for the new field.

▶ Type workingDirectory\car_sales_transformations.xml, substituting the path to the file for workingDirectory, as the location of the XML file containing the transformations.

▶ Click OK.

These selections generate the following command syntax:

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

■ The TMS IMPORT command reads the transformations in *car_sales_transformations.xml* and applies the backtransformation to *PRE_2*.

■ The new field containing the backtransformed values is named *PRE_2_backtransformed*.

■ The EXECUTE command causes the transformations to be processed. When using this as part of a longer stream of syntax, you may be able to remove the EXECUTE command to save some processing time.

Summary

Using automated data preparation, you can quickly obtain transformations of the data that can improve your model. If the target is transformed, you can save the transformations to an XML file and use the Backtransform Scores dialog to convert predicted values for the transformed target back to the original scale.

Identify Unusual Cases

The Anomaly Detection procedure searches for unusual cases based on deviations from the norms of their cluster groups. The procedure is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm is designed for generic anomaly detection; that is, the definition of an anomalous case is not specific to any particular application, such as detection of unusual payment patterns in the healthcare industry or detection of money laundering in the finance industry, in which the definition of an anomaly can be well-defined.

Identify Unusual Cases Algorithm

This algorithm is divided into three stages:

Modeling. The procedure creates a clustering model that explains natural groupings (or clusters) within a dataset that would otherwise not be apparent. The clustering is based on a set of input variables. The resulting clustering model and sufficient statistics for calculating the cluster group norms are stored for later use.

Scoring. The model is applied to each case to identify its cluster group, and some indices are created for each case to measure the unusualness of the case with respect to its cluster group. All cases are sorted by the values of the anomaly indices. The top portion of the case list is identified as the set of anomalies.

Reasoning. For each anomalous case, the variables are sorted by their corresponding variable deviation indices. The top variables, their values, and the corresponding norm values are presented as the reasons why a case is identified as an anomaly.

Identifying Unusual Cases in a Medical Database

A data analyst hired to build predictive models for stroke treatment outcomes is concerned about data quality because such models can be sensitive to unusual observations. Some of these outlying observations represent truly unique cases and are thus unsuitable for prediction, while other observations are caused by data entry errors in which the values are technically "correct" and thus cannot be caught by data validation procedures.

This information is collected in *stroke_valid.sav*. For more information, see the topic Sample Files in Appendix A on p. 134. Use the Identify Unusual Cases procedure to clean the data file. Syntax for reproducing these analyses can be found in *detectanomaly_stroke.sps*.

Running the Analysis

▶ To identify unusual cases, from the menus choose:
Data > Identify Unusual Cases...

Figure 9-1
Identify Unusual Cases dialog box, Variables tab
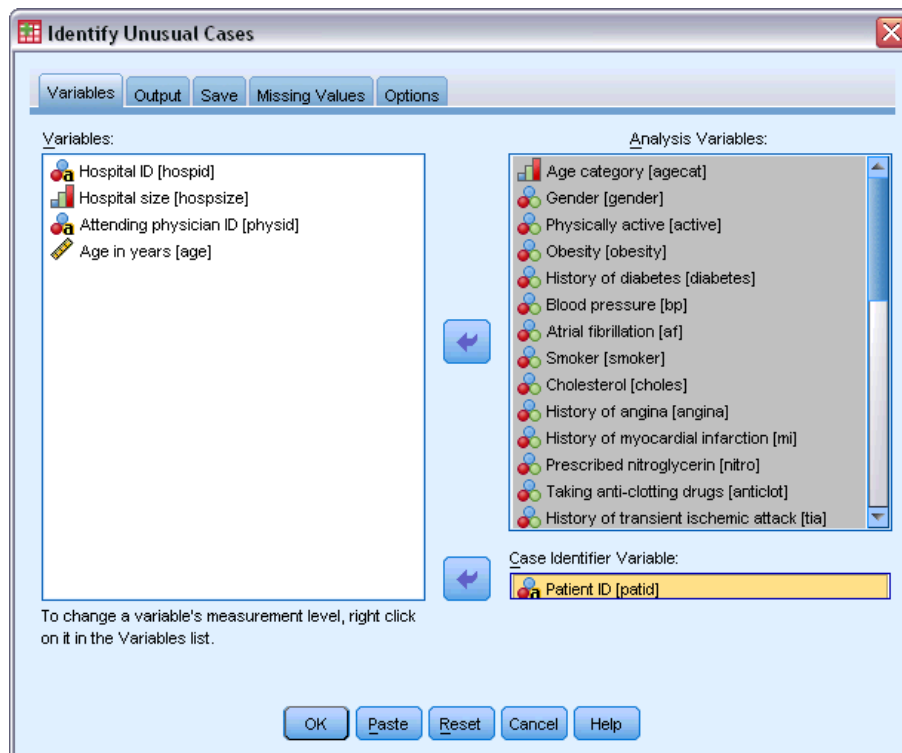
▶ Select *Age category* through *Stroke between 3 and 6 months* as analysis variables.

▶ Select *Patient ID* as the case identifier variable.

▶ Click the Output tab.

Figure 9-2
Identify Unusual Cases dialog box, Output tab
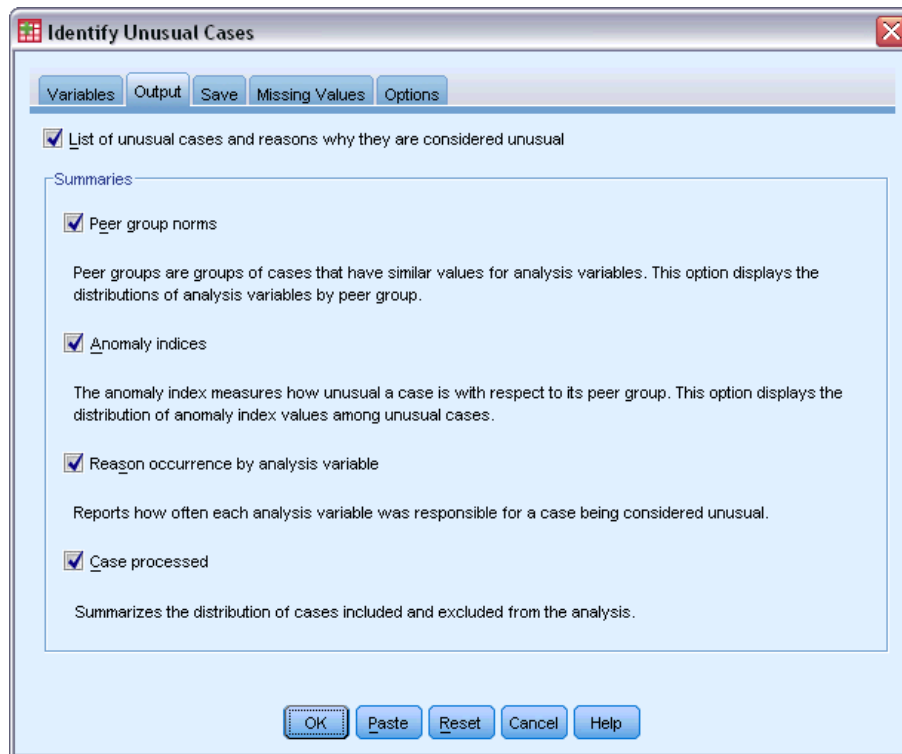
▶ Select Peer group norms, Anomaly indices, Reason occurrence by analysis variable, and Cases processed.

▶ Click the Save tab.

Figure 9-3
Identify Unusual Cases dialog box, Save tab
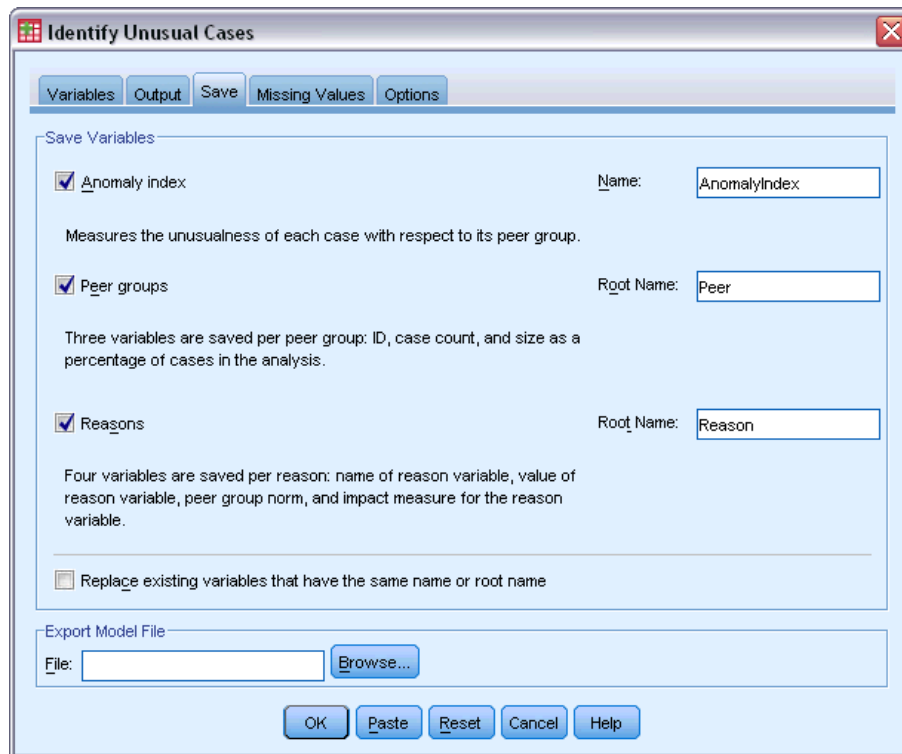
▶ Select Anomaly index, Peer groups, and Reasons.

Saving these results allows you to produce a useful scatterplot that summarizes the results.

▶ Click the Missing Values tab.

Figure 9-4
Identify Unusual Cases dialog box, Missing Values tab
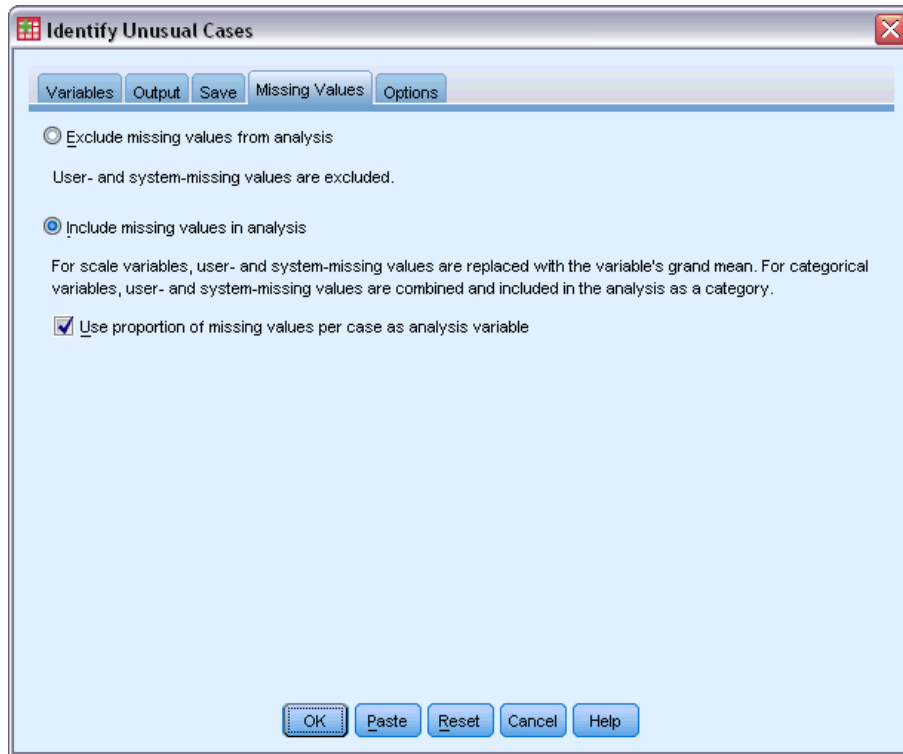
▶ Select Include missing values in analysis. This process is necessary because there are a lot of user-missing values to handle patients who died before or during treatment. An extra variable that measures the proportion of missing values per case is added to the analysis as a scale variable.

▶ Click the Options tab.

Figure 9-5
Identify Unusual Cases dialog box, Options tab

▶ Type 2 as the percentage of cases to consider anomalous.

▶ Deselect Identify only cases whose anomaly index value meets or exceeds a minimum value.

▶ Type 3 as the maximum number of reasons.

▶ Click OK.

Case Processing Summary

Figure 9-6
Case processing summary

		N	% of Combined	% of Total
Peer ID	1	710	67.7%	67.7%
	2	90	8.6%	8.6%
	3	248	23.7%	23.7%
Combined		1048	100.0%	100.0%
Total		1048		100.0%

Each case is categorized into a peer group of similar cases. The case processing summary shows the number of peer groups that were created, as well as the number and percentage of cases in each peer group.

Anomaly Case Index List

Figure 9-7
Anomaly case index list

Case	patid	Anomaly Index
843	7840326167	2.837
510	0714726620	2.022
623	6553808330	2.014
501	6461046805	2.002
607	1077125669	1.897
884	2260043998	1.889
614	4030164769	1.869
241	1038840465	1.865
13	2191527525	1.826
172	4458028382	1.786
705	1336411777	1.778
651	4103977868	1.767
384	2247641363	1.767
839	0437454972	1.766
861	9746101913	1.757
19	7237535360	1.756
806	4391632997	1.756
871	6961938294	1.739
239	7315965190	1.738
887	6044244232	1.737
245	0816869249	1.736

The anomaly index is a measure that reflects the unusualness of a case with respect to its peer group. The 2% of cases with the highest values of the anomaly index are displayed, along with their case numbers and IDs. Twenty-one cases are listed, ranging in value from 1.736 to 2.837. There is a relatively large difference in the value of the anomaly index between the first and second cases in the list, which suggests that case 843 is probably anomalous. The other cases will need to be judged on a case-by-case basis.

Anomaly Case Peer ID List

Figure 9-8
Anomaly case peer ID list

Case	patid	Peer ID	Peer Size	Peer Size Percent
843	7840326167	3	248	23.7%
510	0714726620	3	248	23.7%
623	6553808330	3	248	23.7%
501	6461046805	3	248	23.7%
607	1077125669	3	248	23.7%
884	2260043998	3	248	23.7%
614	4030164769	3	248	23.7%
241	1038840465	3	248	23.7%
13	2191527525	3	248	23.7%
172	4458028382	3	248	23.7%
705	1336411777	1	710	67.7%
651	4103977868	1	710	67.7%
384	2247641363	3	248	23.7%
839	0437454972	3	248	23.7%
861	9746101913	3	248	23.7%
19	7237535360	1	710	67.7%
806	4391632997	1	710	67.7%
871	6961938294	1	710	67.7%
239	7315965190	3	248	23.7%
887	6044244232	1	710	67.7%
245	0816869249	3	248	23.7%

The potentially anomalous cases are displayed with their peer group membership information. The first 10 cases, and 15 cases overall, belong to peer group 3, with the remainder belonging to peer group 1.

Anomaly Case Reason List

Figure 9-9
Anomaly case reason list

Reason: 1

Case	patid	Reason Variable	Variable Impact	Variable Value	Variable Norm
843	7840326167	cost	.411	200.51	19.83
510	0714726620	cost	.120	96.59	19.83
623	6553808330	cost	.175	114.01	19.83
501	6461046805	barthel1	.084	80	(Missing Value)
607	1077125669	cost	.126	96.11	19.83
884	2260043998	cost	.138	99.73	19.83
614	4030164769	barthel1	.085	45	(Missing Value)
241	1038840465	barthel1	.115	25	(Missing Value)
13	2191527525	barthel1	.118	40	(Missing Value)
172	4458028382	barthel1	.120	100	(Missing Value)
705	1336411777	cost	.244	198.25	42.47
651	4103977868	barthel1	.064	30	95
384	2247641363	barthel1	.122	20	(Missing Value)
839	0437454972	barthel1	.109	95	(Missing Value)
861	9746101913	barthel1	.102	70	(Missing Value)
19	7237535360	barthel3	.080	5	100
806	4391632997	barthel2	.088	10	100
871	6961938294	barthel1	.094	5	95
239	7315965190	barthel1	.092	45	(Missing Value)
887	6044244232	barthel1	.066	40	95
245	0816869249	barthel1	.124	5	(Missing Value)

Reason variables are the variables that contribute the most to a case's classification as unusual. The primary reason variable for each anomalous case is displayed, along with its impact, value for that case, and peer group norm. The peer group norm *(Missing Value)* for a categorical variable indicates that the plurality of cases in the peer group had a missing value for the variable.

The variable impact statistic is the proportional contribution of the reason variable to the deviation of the case from its peer group. With 38 variables in the analysis, including the missing proportion variable, a variable's expected impact would be $1/38 = 0.026$. The impact of the variable *cost* on case 843 is 0.411, which is relatively large. The value of *cost* for case 843 is 200.51, compared to the average of 19.83 for cases in peer group 3.

The dialog box selections requested results for the top three reasons.

▶ To see the results for the other reasons, activate the table by double-clicking it.

▶ Move *Reason* from the layer dimension to the row dimension.

Figure 9-10
Anomaly case reason list (first 8 cases)

Case	Reason	patid	Reason Variable	Variable Impact	Variable Value	Variable Norm
843	1	7840326167	cost	.411	200.51	19.83
	2	7840326167	barthel1	.076	65	(Missing Value)
	3	7840326167	rankin1	.044	2	(Missing Value)
510	1	0714726620	cost	.120	96.59	19.83
	2	0714726620	barthel1	.083	80	(Missing Value)
	3	0714726620	rehab	.068	3	(Missing Value)
623	1	6553808330	cost	.175	114.01	19.83
	2	6553808330	surgery	.089	2	(Missing Value)
	3	6553808330	barthel1	.089	70	(Missing Value)
501	1	6461046805	barthel1	.084	80	(Missing Value)
	2	6461046805	rehab	.068	3	(Missing Value)
	3	6461046805	rankin1	.063	1	(Missing Value)
607	1	1077125669	cost	.126	96.11	19.83
	2	1077125669	barthel1	.094	85	(Missing Value)
	3	1077125669	rehab	.072	3	(Missing Value)
884	1	2260043998	cost	.138	99.73	19.83
	2	2260043998	barthel1	.114	65	(Missing Value)
	3	2260043998	rehab	.072	3	(Missing Value)
614	1	4030164769	barthel1	.085	45	(Missing Value)
	2	4030164769	rankin1	.085	3	(Missing Value)
	3	4030164769	recbart1	.062	2	(Missing Value)

This configuration makes it easy to compare the relative contributions of the top three reasons for each case. Case 843 is, as suspected, considered anomalous because of its unusually large value of *cost*. In contrast, no single reason contributes more than 0.10 to the unusualness of case 501.

Scale Variable Norms

Figure 9-11
Scale variable norms

		Peer ID			
		1	2	3	Combined
Length of stay for rehabilitation	Mean	16.55	16.39	15.91	16.39
	Std. Deviation	12.596	.000	6.834	10.887
Total treatment and rehabilitation costs in thousands	Mean	42.4673	3.5089	19.8273	33.7641
	Std. Deviation	26.45401	.50997	20.17309	27.31266
Missing Proportion	Mean	.006	.541	.354	.134
	Std. Deviation	.021	2.9E-016	.083	.197

The scale variable norms report the mean and standard deviation of each variable for each peer group and overall. Comparing the values gives some indication of which variables contribute to peer group formation.

For example, the mean for *Length of stay for rehabilitation* is fairly constant across all three peer groups, meaning that this variable does not contribute to peer group formation. In contrast, *Total treatment and rehabilitation costs in thousands* and *Missing Proportion* each provide some insight into peer group membership. Peer group 1 has the highest average cost and the fewest

missing values. Peer group 2 has very low costs and a lot of missing values. Peer group 3 has middling costs and missing values.

This organization suggests that peer group 2 is composed of patients who were dead on arrival, thus incurring very little cost and causing all of the treatment and rehabilitation variables to be missing. Peer group 3 likely contains many patients who died during treatment, thus incurring the treatment costs but not the rehabilitation costs and causing the rehabilitation variables to be missing. Peer group 1 is likely composed almost entirely of patients who survived through treatment and rehabilitation, thus incurring the highest costs.

Categorical Variable Norms

Figure 9-12
Categorical variable norms (first 10 variables)

		Peer ID			
		1	2	3	Combined
Age category	Most Popular Category	2	3	2	2
	Frequency	277	25	81	383
	Percent	39.0%	27.8%	32.7%	36.5%
Gender	Most Popular Category	0	0	1	0
	Frequency	361	46	126	529
	Percent	50.8%	51.1%	50.8%	50.5%
Physically active	Most Popular Category	1	0	0	0
	Frequency	373	55	139	531
	Percent	52.5%	61.1%	56.0%	50.7%
Obesity	Most Popular Category	0	0	0	0
	Frequency	555	67	178	800
	Percent	78.2%	74.4%	71.8%	76.3%
History of diabetes	Most Popular Category	0	0	0	0
	Frequency	665	80	219	964
	Percent	93.7%	88.9%	88.3%	92.0%
Blood pressure	Most Popular Category	1	1	1	1
	Frequency	445	49	139	633
	Percent	62.7%	54.4%	56.0%	60.4%
Atrial fibrillation	Most Popular Category	0	0	0	0
	Frequency	641	83	216	940
	Percent	90.3%	92.2%	87.1%	89.7%
Smoker	Most Popular Category	0	0	0	0
	Frequency	578	69	179	826
	Percent	81.4%	76.7%	72.2%	78.8%
Cholesterol	Most Popular Category	0	0	0	0
	Frequency	406	52	136	594
	Percent	57.2%	57.8%	54.8%	56.7%
History of angina	Most Popular Category	0	0	0	0
	Frequency	493	52	167	712
	Percent	69.4%	57.8%	67.3%	67.9%

The categorical variable norms serve much the same purpose as the scale norms, but categorical variable norms report the modal (most popular) category and the number and percentage of cases in the peer group that fall into that category. Comparing the values can be somewhat trickier; for example, at first glance, it may appear that *Gender* contributes more to cluster formation than *Smoker* because the modal category for *Smoker* is the same for all three peer groups, while the modal category for *Gender* differs on peer group 3. However, because *Gender* has only two values, you can infer that 49.2% of the cases in peer group 3 have a value of 0, which is very

similar to the percentages in the other peer groups. By contrast, the percentages for *Smoker* range from 72.2% to 81.4%.

Figure 9-13
Categorical variable norms (selected variables)

		Peer ID			
		1	2	3	Combined
Dead on arrival	Most Popular Category	0	1	0	0
	Frequency	710	90	248	958
	Percent	100.0%	100.0%	100.0%	91.4%
Initial Rankin score	Most Popular Category	0	(Missing Value)	5	5
	Frequency	166	90	104	193
	Percent	23.4%	100.0%	41.9%	18.4%
CAT scan result	Most Popular Category	0	(Missing Value)	0	0
	Frequency	607	90	184	791
	Percent	85.5%	100.0%	74.2%	75.5%
Clot-dissolving drugs	Most Popular Category	2	(Missing Value)	0	2
	Frequency	318	90	129	394
	Percent	44.8%	100.0%	52.0%	37.6%
Died in hospital	Most Popular Category	0	(Missing Value)	1	0
	Frequency	710	90	171	787
	Percent	100.0%	100.0%	69.0%	75.1%
Treatment result	Most Popular Category	1	(Missing Value)	1	1
	Frequency	524	90	96	620
	Percent	73.8%	100.0%	38.7%	59.2%
Post-event preventative surgery	Most Popular Category	0	(Missing Value)	(Missing Value)	0
	Frequency	323	90	171	369
	Percent	45.5%	100.0%	69.0%	35.2%
Post-event rehabilitation	Most Popular Category	0	(Missing Value)	(Missing Value)	0
	Frequency	278	90	171	314
	Percent	39.2%	100.0%	69.0%	30.0%

The suspicions that were raised by the scale variable norms are confirmed further down in the categorical norms table. Peer group 2 is composed entirely of patients who were dead on arrival, so all treatment and rehabilitation variables are missing. Most of the patients in peer group 3 (69.0%) died during treatment, so the modal category for rehabilitation variables is *(Missing Value)*.

Anomaly Index Summary

Figure 9-14
Anomaly index summary

	N in the Anomaly List	Minimum	Maximum	Mean	Std. Deviation
Anomaly Index	21	1.736	2.837	1.872	.240

N in the Anomaly List is determined by the specification: anomaly percentage is 2%

The table provides summary statistics for the anomaly index values of cases in the anomaly list.

Reason Summary

Figure 9-15
Reason summary (treatment and rehabilitation variables)

	Occurrence as Reason		Variable Impact Statistics			
	Frequency	Percent	Minimum	Maximum	Mean	Std. Deviation
Dead on arrival	0	.0%	.	.	.	.
Initial Rankin score	0	.0%	.	.	.	.
CAT scan result	0	.0%	.	.	.	.
Clot-dissolving drugs	0	.0%	.	.	.	.
Died in hospital	0	.0%	.	.	.	.
Treatment result	0	.0%	.	.	.	.
Post-event preventative surgery	0	.0%	.	.	.	.
Post-event rehabilitation	0	.0%	.	.	.	.
Rankin score at 1 month	0	.0%	.	.	.	.
Rankin score at 3 months	0	.0%	.	.	.	.
Rankin score at 6 months	0	.0%	.	.	.	.
Barthel index at 1 month	13	61.9%	.064	.124	.100	.021
Barthel index at 3 months	1	4.8%	.088	.088	.088	.
Barthel index at 6 months	1	4.8%	.080	.080	.080	.
Recoded Barthel index at 1 month	0	.0%	.	.	.	.
Recoded Barthel index at 3 months	0	.0%	.	.	.	.
Recoded Barthel index at 6 months	0	.0%	.	.	.	.
Stroke between release and 1 month	0	.0%	.	.	.	.
Stroke between 1 and 3 months	0	.0%	.	.	.	.
Stroke between 3 and 6 months	0	.0%	.	.	.	.
Length of stay for rehabilitation	0	.0%	.	.	.	.
Total treatment and rehabilitation costs in thousands	6	28.6%	.120	.411	.202	.112
Missing Proportion	0	.0%	.	.	.	.
Overall	21	100.0%	.064	.411	.127	.076

For each variable in the analysis, the table summarizes the variable's role as a primary reason. Most variables, such as variables from *Dead on arrival* to *Post-event rehabilitation*, are not the primary reason that any of the cases are on the anomaly list. *Barthel index at 1 month* is the most frequent reason, followed by *Total treatment and rehabilitation costs in thousands*. The variable impact statistics are summarized, with the minimum, maximum, and mean impact reported for each variable, along with the standard deviation for variables that were the reason for more than one case.

Scatterplot of Anomaly Index by Variable Impact

The tables contain a lot of useful information, but it can be difficult to grasp the relationships. Using the saved variables, you can construct a graph that makes this process easier.

▶ To produce this scatterplot, from the menus choose:
Graphs > Chart Builder...

Figure 9-16
Chart Builder dialog box
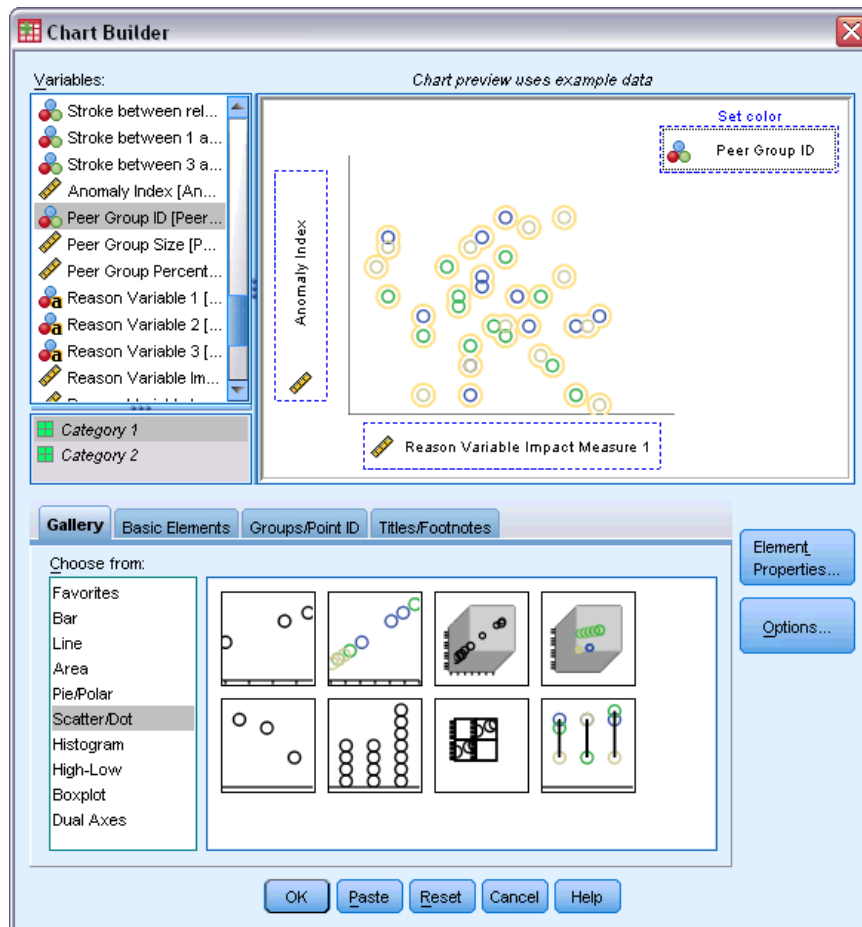
▶ Select the Scatter/Dot gallery and drag the Grouped Scatter icon onto the canvas.

▶ Select *Anomaly Index* as the *y* variable and *Reason Variable Impact Measure 1* as the *x* variable.

▶ Select *Peer Group ID* as the variable to set colors by.

▶ Click OK.

These selections produce the scatterplot.

Figure 9-17
Scatterplot of anomaly index by impact measure of first reason variable
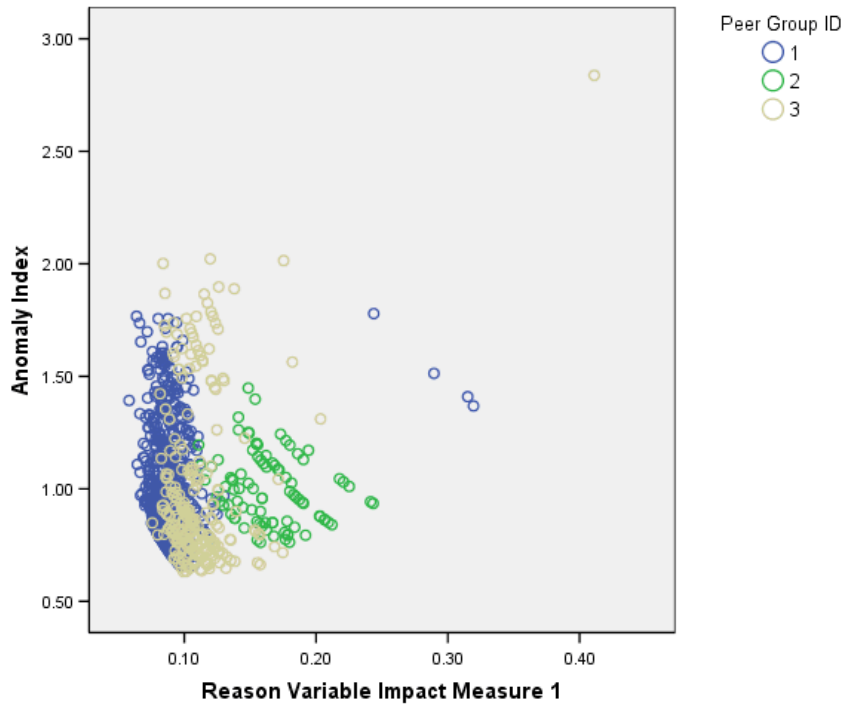
Inspection of the graph leads to several observations:

■ The case in the upper right corner belongs to peer group 3 and is both the most anomalous case and the case with the largest contribution made by a single variable.

■ Moving down along the *y* axis, we see that there are three cases belonging to peer group 3, with anomaly index values just above 2.00. These cases should be investigated more closely as anomalous.

■ Moving along the *x* axis, we see that there are four cases belonging to peer group 1, with variable impact measures approximately in the range of 0.23 to 0.33. These cases should be investigated more thoroughly because these values separate the cases from the main body of points in the plot.

■ Peer group 2 seems fairly homogenous in the sense that its anomaly index and variable impact values do not vary widely from their central tendencies.

Summary

Using the Identify Unusual Cases procedure, you have spotted several cases that warrant further examination. These cases are ones that would not be identified by other validation procedures, because the relationships between the variables (not just the values of the variables themselves) determine the anomalous cases.

It is somewhat disappointing that the peer groups are largely constructed based on two variables: *Dead on arrival* and *Died in hospital*. In further analysis, you could study the effect of forcing a larger number of peer groups to be created, or you could perform an analysis that includes only patients who have survived treatment.

Related Procedures

The Identify Unusual Cases procedure is a useful tool for detecting anomalous cases in your data file.

■ The Validate Data procedure identifies suspicious and invalid cases, variables, and data values in the active dataset.

Optimal Binning

The Optimal Binning procedure discretizes one or more scale variables (referred to as **binning input variables**) by distributing the values of each variable into bins. Bin formation is optimal with respect to a categorical guide variable that "supervises" the binning process. Bins can then be used instead of the original data values for further analysis in procedures that require or prefer categorical variables.

The Optimal Binning Algorithm

The basic steps of the Optimal Binning algorithm can be characterized as follows:

Preprocessing (optional). The binning input variable is divided into *n* bins (where *n* is specified by you), and each bin contains the same number of cases or as near the same number of cases as possible.

Identifying potential cutpoints. Each distinct value of the binning input that does not belong to the same category of the guide variable as the next larger distinct value of the binning input variable is a potential cutpoint.

Selecting cutpoints. The potential cutpoint that produces the greatest information gain is evaluated by the MDLP acceptance criterion. Repeat until no potential cutpoints are accepted. The accepted cutpoints define the endpoints of the bins.

Using Optimal Binning to Discretize Loan Applicant Data

As part of a bank's efforts to reduce the rate of loan defaults, a loan officer has collected financial and demographic information on past and present customers in the hopes of creating a model for predicting the probability of loan default. Several potential predictors are scale, but the loan officer wants to be able to consider models that work best with categorical predictors.

Information on 5000 past customers is collected in *bankloan_binning.sav*. For more information, see the topic Sample Files in Appendix A on p. 134. Use the Optimal Binning procedure to generate binning rules for the scale predictors, and then use the generated rules to process *bankloan.sav*. The processed dataset can then be used to create a predictive model.

Running the Analysis

▶ To run an Optimal Binning analysis, from the menus choose:
Transform > Optimal Binning...

Figure 10-1
Optimal Binning dialog box, Variables tab
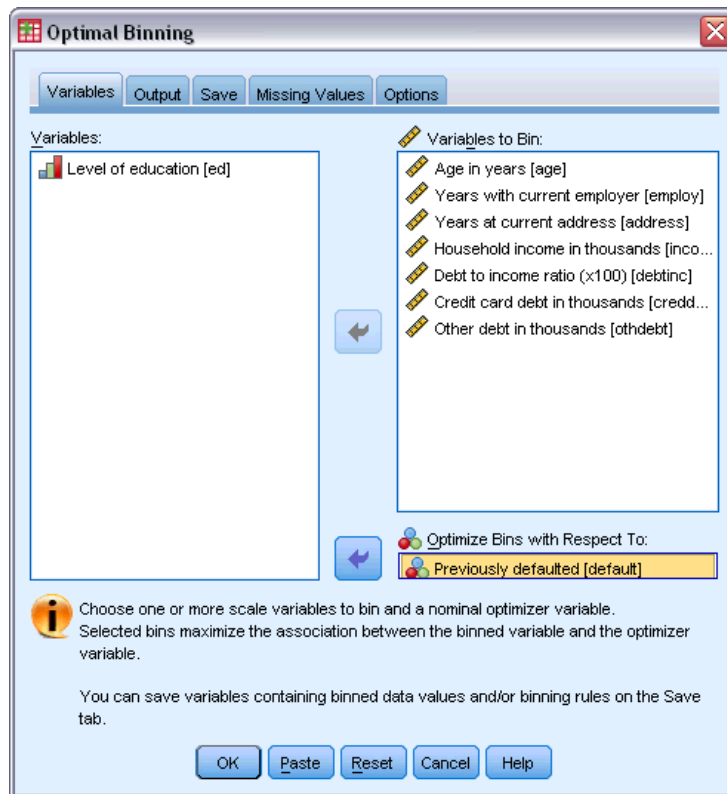
▶ Select *Age in years* and *Years with current employer* through *Other debt in thousands* as variables to bin.

▶ Select *Previously defaulted* as the guide variable.

▶ Click the Output tab.

Figure 10-2
Optimal Binning dialog box, Output tab
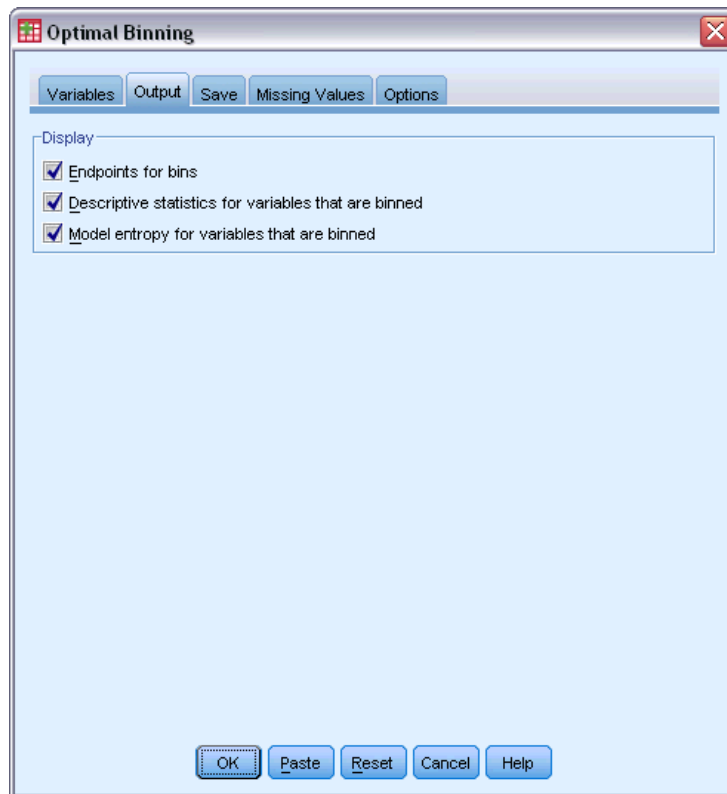
▶ Select Descriptive statistics and Model entropy for variables that are binned.

▶ Click the Save tab.

Figure 10-3
Optimal Binning dialog box, Save tab
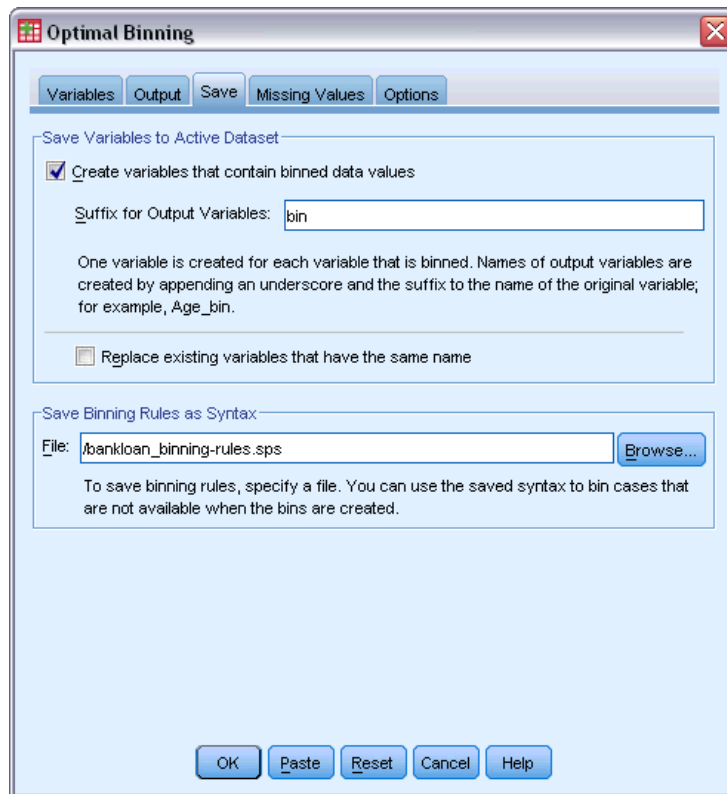
▶ Select Create variables that contain binned data values.

▶ Enter a path and filename for the syntax file to contain the generated binning rules. In this example, we have used */bankloan_binning-rules.sps*.

▶ Click OK.

These selections generate the following command syntax:

```
* Optimal Binning.
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
  othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bin
  creddebt_bin othdebt_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
  /OUTFILE RULES='/bankloan_binning-rules.sps'
  /PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

■ The procedure will discretize the binning input variables *age*, *employ*, *address*, *income*, *debtinc*, *creddebt*, and *othdebt* using MDLP binning with the guide variable *default*.

- The discretized values for these variables will be stored in the new variables *age_bin*, *employ_bin*, *address_bin*, *income_bin*, *debtinc_bin*, *creddebt_bin*, and *othdebt_bin*.

- If a binning input variable has more than 1000 distinct values, then the equal frequency method will reduce the number to 1000 before performing MDLP binning.

- Command syntax representing the binning rules is saved to the file */bankloan_binning-rules.sps*.

- Bin endpoints, descriptive statistics, and model entropy values are requested for binning input variables.

- Other binning criteria are set to their default values.

Descriptive Statistics

Figure 10-4
Descriptive statistics

	N	Minimum	Maximum	Number of Distinct Values	Number of Bins
Age in years	5000	20	58	39	2
Years with current employer	5000	0	38	39	4
Years at current address	5000	0	37	38	3
Household income in thousands	5000	12.10	2461.70	1100	2
Debt to income ratio (x100)	5000	.08	44.62	2060	5
Credit card debt in thousands	5000	.01	139.58	5000	4
Other debt in thousands	5000	.01	416.52	4999	2

The descriptive statistics table provides summary information on the binning input variables. The first four columns concern the pre-binned values.

- N is the number of cases used in the analysis. When listwise deletion of missing values is used, this value should be constant across variables. When pairwise missing value handling is used, this value may not be constant. Since this dataset has no missing values, the value is simply the number of cases.

- The Minimum and Maximum columns show the (pre-binning) minimum and maximum values in the dataset for each binning input variable. In addition to giving a sense of the observed range of values for each variable, these can be useful for catching values outside the expected range.

- The Number of Distinct Values tells you which variables were preprocessed using the equal frequencies algorithm. By default, variables with more than 1000 distinct values (*Household income in thousands* through *Other debt in thousands*) are pre-binned into 1000 distinct bins. These preprocessed bins are then binned against the guide variable using MDLP. You can control the preprocessing feature on the Options tab.

- The Number of Bins is the final number of bins generated by the procedure and is much smaller than the number of distinct values.

Model Entropy

Figure 10-5
Model entropy

	Model Entropy
Age in years	.788
Years with current employer	.754
Years at current address	.781
Household income in thousands	.803
Debt to income ratio (x100)	.711
Credit card debt in thousands	.776
Other debt in thousands	.801

Smaller model entropy indicates higher predictive accuracy of
the binned variable on guide variable Previously defaulted.

The model entropy gives you an idea of how useful each variable could be in a predictive model for the probability of default.

■ The best possible predictor is one that, for each generated bin, contains cases with the same value as the guide variable; thus, the guide variable can be perfectly predicted. Such a predictor has an undefined model entropy. This generally does not occur in real-world situations and may indicate problems with the quality of your data.

■ The worst possible predictor is one that does no better than guessing; the value of its model entropy is dependent upon the data. In this dataset, 1256 (or 0.2512) of the 5000 total customers defaulted and 3744 (or 0.7488) did not; thus, the worst possible predictor would have a model entropy of $-0.2512 \times \log_2(0.2512) - 0.7488 \times \log_2(0.7488) = 0.8132$.

It is difficult to make a statement more conclusive than that variables with lower model entropy values should make better predictors, since what constitutes a good model entropy value is application and data-dependent. In this case, it appears that variables with a larger number of generated bins, relative to the number of distinct categories, have lower model entropy values. Further evaluation of these binning input variables as predictors should be performed using predictive modeling procedures, which have more extensive tools for variable selection.

Binning Summaries

The binning summary reports the bounds of the generated bins and the frequency count of each bin by values of the guide variable. A separate binning summary table is produced for each binning input variable.

Figure 10-6
Binning summary for Age in years

	End Point		Number of Cases by Level of Previously defaulted		
Bin	Lower	Upper	No	Yes	Total
1	a	32	1129	639	1768
2	32	a	2615	617	3232
Total			3744	1256	5000

Each bin is computed as Lower <= Age in years < Upper.
  a. Unbounded

128

Chapter 10

The summary for *Age in years* shows that 1768 customers, all aged 32 years or younger, are put into Bin 1, while the remaining 3232 customers, all greater than 32 years of age, are put into Bin 2. The proportion of customers who previously defaulted is much higher in Bin 1 (639/1768=0.361) than in Bin 2 (617/3232=0.191).

Figure 10-7
Binning summary for Household income in thousands

| Bin | End Point | | Number of Cases by Level of Previously defaulted | | |
	Lower	Upper	No	Yes	Total
1	a	26.70	1054	513	1567
2	26.70	a	2690	743	3433
Total			3744	1256	5000

Each bin is computed as Lower <= Household income in thousands < Upper.
a. Unbounded

The summary for *Household income in thousands* shows a similar pattern, with a single cutpoint at 26.70 and a higher proportion of customers who previously defaulted in Bin 1 (513/1567=0.327) than in Bin 2 (743/3433=0.216). As expected from the model entropy statistics, the difference in these proportions is not as great as that for *Age in years*.

Figure 10-8
Binning summary for Other debt in thousands

| Bin | End Point | | Number of Cases by Level of Previously defaulted | | |
	Lower	Upper	No	Yes	Total
1	a	2.19	2161	539	2700
2	2.19	a	1583	717	2300
Total			3744	1256	5000

Each bin is computed as Lower <= Other debt in thousands < Upper.
a. Unbounded

The summary for *Other debt in thousands* shows a reversed pattern, with a single cutpoint at 2.19 and a lower proportion of customers who previously defaulted in Bin 1 (539/2700=0.200) than in Bin 2 (717/2300=0.312). Again, as expected from the model entropy statistics, the difference in these proportions is not as great as that for *Age in years*.

Figure 10-9
Binning summary for Years with current employer

Bin	End Point Lower	End Point Upper	No	Yes	Total
1	a	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	a	578	49	627
Total			3744	1256	5000

Number of Cases by Level of Previously defaulted

Each bin is computed as Lower <= Years with current employer < Upper.
a. Unbounded

The summary for *Years with current employer* shows a pattern of decreasing proportions of defaulters as the bin numbers increase.

Bin	Proportion of Defaulters
1	0.432
2	0.302
3	0.154
4	0.078

Figure 10-10
Binning summary for Years at current address

Bin	End Point Lower	End Point Upper	No	Yes	Total
1	a	7	1652	829	2481
2	7	14	1184	313	1497
3	14	a	908	114	1022
Total			3744	1256	5000

Number of Cases by Level of Previously defaulted

Each bin is computed as Lower <= Years at current address < Upper.
a. Unbounded

The summary for *Years at current address* shows a similar pattern. As expected from the model entropy statistics, the differences between bins in the proportion of defaulters is sharper for *Years with current employer* than *Years at current address*.

Bin	Proportion of Defaulters
1	0.334
2	0.209
3	0.112

Figure 10-11
Binning summary for Credit card debt in thousands

Bin	End Point		Number of Cases by Level of Previously defaulted		
	Lower	Upper	No	Yes	Total
1	a	.97	2169	466	2635
2	.97	1.91	848	307	1155
3	1.91	6.05	643	352	995
4	6.05	a	84	131	215
Total			3744	1256	5000

Each bin is computed as Lower <= Credit card debt in thousands < Upper.

a. Unbounded

The summary for *Credit card debt in thousands* shows the reverse pattern, with increasing proportions of defaulters as the bin numbers increase. *Years with current employer* and *Years at current address* seem to do a better job of identifying high-probability nondefaulters, while *Credit card debt in thousands* does a better job of identifying high-probability defaulters.

Bin	Proportion of Defaulters
1	0.177
2	0.266
3	0.354
4	0.609

Figure 10-12
Binning summary for Debt to income ratio (x100)

Bin	End Point		Number of Cases by Level of Previously defaulted		
	Lower	Upper	No	Yes	Total
1	a	4.39	912	88	1000
2	4.39	12.09	2006	437	2443
3	12.09	18.71	625	386	1011
4	18.71	31.00	198	303	501
5	31.00	a	3	42	45
Total			3744	1256	5000

Each bin is computed as Lower <= Debt to income ratio (x100) < Upper.

a. Unbounded

The summary for *Debt to income ratio (x100)* shows a similar pattern to *Credit card debt in thousands*. This variable has the lowest model entropy value and is thus the best prospective predictor of the probability of default. It does a better job of classifying high-probability defaulters than *Credit card debt in thousands* and almost as good of a job of classifying low-probability defaulters as *Years with current employer*.

Bin	Proportion of Defaulters
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

Binned Variables

Figure 10-13
Binned variables for bankloan_binning.sav in Data Editor

	default	age_bin	employ_bin	address_bi	income_bin	debtinc_bin	creddebt_bi	othdebt_bin
1	0	2	3	2	2	2	1	2
2	0	1	3	2	2	3	2	2
3	0	2	3	3	2	2	3	2
4	0	2	3	3	2	4	3	2
5	0	2	2	3	1	3	2	2
6	0	2	1	2	2	1	1	1
7	1	2	1	1	1	3	2	1
8	0	2	4	2	2	3	2	2
9	0	2	3	2	2	2	2	2
10	0	2	2	2	2	2	2	2
11	0	1	1	1	1	2	1	1
12	1	2	3	2	2	4	4	2
13	0	2	3	3	2	2	3	2
14	1	2	3	1	2	2	1	1
15	0	1	1	2	2	2	2	1

Data View / Variable View

The results of the binning process on this dataset are evident in the Data Editor. These binned variables are useful if you want to produce customized summaries of the binning results using descriptive or reporting procedures, but it is inadvisable to use this dataset to build a predictive model because the binning rules were generated using these cases. A better plan is to apply the binning rules to another dataset containing information on other customers.

Applying Syntax Binning Rules

While running the Optimal Binning procedure, you requested that the binning rules generated by the procedure be saved as command syntax.

▶ Open *bankloan_binning-rules.sps*.

Figure 10-14
Syntax rules file

For each binning input variable, there is a block of command syntax that performs the binning; sets the variable label, format, and level; and sets the value labels for the bins. These commands can be applied to a dataset with the same variables as *bankloan_binning.sav*.

► Open *bankloan.sav*. For more information, see the topic Sample Files in Appendix A on p. 134.

► Return to the Syntax Editor view of *bankloan_binning-rules.sps*.

► To apply the binning rules, from the Syntax Editor menus choose:
Run > All...

Figure 10-15
Binned variables for bankloan.sav in Data Editor

The variables in *bankloan.sav* have been binned according to the rules generated by running the Optimal Binning procedure on *bankloan_binning.sav*. This dataset is now ready for use in building predictive models that prefer or require categorical variables.

Summary

Using the Optimal Binning procedure, we have generated binning rules for scale variables that are potential predictors for the probability of default and applied these rules to a separate dataset.

During the binning process, you noted that the binned *Years with current employer* and *Years at current address* seem to do a better job of identifying high-probability non-defaulters, while the *Credit card debt in thousands* does a better job of identifying high-probability defaulters. This interesting observation will give you some extra insight when building predictive models for the probability of default. If avoiding bad debt is a primary concern, then *Credit card debt in thousands* will be more important than *Years with current employer* and *Years at current address*. If growing your customer base is the priority, then *Years with current employer* and *Years at current address* will be more important.

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 (Van der Ham, Meulman, Van Strien, and Van Engeland, 1997) 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.

- **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.

- **behavior.sav.** In a classic example (Price and Bouffard, 1974), 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 (Green and Rao, 1972), 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.

- **car_insurance_claims.sav.** A dataset presented and analyzed elsewhere (McCullagh and Nelder, 1989) 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 (Green and Wind, 1973), 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 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 (Kennedy, Riquier, and Sharp, 1996) . 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.

- **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.

- **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.

- **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.

- **dietstudy.sav.** This hypothetical data file contains the results of a study of the "Stillman diet" (Rickman, Mitchell, Dingman, and Dalen, 1974). 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.

- **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.

- **german_credit.sav.** This data file is taken from the "German credit" dataset in the Repository of Machine Learning Databases (Blake and Merz, 1998) 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.

- **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 (Bell, 1961) presented a table to illustrate possible social groups. Guttman (Guttman, 1968) 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 (Rosenberg and Kim, 1975) 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 15×15 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 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 (Breiman and Friedman, 1985), (Hastie and Tibshirani, 1990), 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 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 (Hartigan, 1975).

- **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 (McCullagh et al., 1989) 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 the Data Preparation option.

- **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.

- **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 (Collett, 2003).

- **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 (Collett et al., 2003).

- **verd1985.sav.** This data file concerns a survey (Verdegaal, 1985). 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 (Ware, Dockery, Spiro III, Speizer, and Ferris Jr., 1984). The data contain repeated binary measures of the wheezing status for children 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 worldwide.

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, Armonk, NY 10504-1785, U.S.A.

For license inquiries regarding double-byte character set (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., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES 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 states 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 Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply 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 Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

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.

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-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

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 the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, ibm.com, and SPSS are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://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.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

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.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Bibliography

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

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

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

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

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

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

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

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

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

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.

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

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.

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.

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.

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.

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

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.

Index