



# > SPSS Data Preparation™ 15.0



For more information about SPSS® software products, please visit our Web site at <http://www.spss.com> or contact

SPSS Inc.

233 South Wacker Drive, 11th Floor

Chicago, IL 60606-6412

Tel: (312) 651-3000

Fax: (312) 651-3668

SPSS is a registered trademark and the other product names are the trademarks of SPSS Inc. for its proprietary computer software. No material describing such software may be produced or distributed without the written permission of the owners of the trademark and license rights in the software and the copyrights in the published materials.

The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c) (1) (ii) of The Rights in Technical Data and Computer Software clause at 52.227-7013. Contractor/manufacturer is SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.

Patent No. 7,023,453

General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

TableLook is a trademark of SPSS Inc.

Windows is a registered trademark of Microsoft Corporation.

DataDirect, DataDirect Connect, INTERSOLV, and SequeLink are registered trademarks of DataDirect Technologies.

Portions of this product were created using LEADTOOLS © 1991–2000, LEAD Technologies, Inc. ALL RIGHTS RESERVED.

LEAD, LEADTOOLS, and LEADVIEW are registered trademarks of LEAD Technologies, Inc.

Sax Basic is a trademark of Sax Software Corporation. Copyright © 1993–2004 by Polar Engineering and Consulting. All rights reserved.

A portion of the SPSS software contains zlib technology. Copyright © 1995–2002 by Jean-loup Gailly and Mark Adler. The zlib software is provided “as is,” without express or implied warranty.

A portion of the SPSS software contains Sun Java Runtime libraries. Copyright © 2003 by Sun Microsystems, Inc. All rights reserved.

The Sun Java Runtime libraries include code licensed from RSA Security, Inc. Some portions of the libraries are licensed from IBM and are available at <http://www-128.ibm.com/developerworks/opensource/>.

SPSS Data Preparation™ 15.0

Copyright © 2006 by SPSS Inc.

All rights reserved.

Printed in the United States of America.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

1 2 3 4 5 6 7 8 9 0 09 08 07 06

ISBN-13: 978-1-56827-385-3

ISBN-10: 1-56827-385-1

---

# ***Preface***

SPSS 15.0 is a comprehensive system for analyzing data. The SPSS 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 15.0 Base system and is completely integrated into that system.

## ***Installation***

To install the SPSS Data Preparation add-on module, run the License Authorization Wizard using the authorization code that you received from SPSS Inc. For more information, see the installation instructions supplied with the SPSS Data Preparation add-on module.

## ***Compatibility***

SPSS is designed to run on many computer systems. See the installation instructions that came with your system for specific information on minimum and recommended requirements.

## ***Serial Numbers***

Your serial number is your identification number with SPSS Inc. You will need this serial number when you contact SPSS Inc. for information regarding support, payment, or an upgraded system. The serial number was provided with your Base system.

## ***Customer Service***

If you have any questions concerning your shipment or account, contact your local office, listed on the SPSS Web site at <http://www.spss.com/worldwide>. Please have your serial number ready for identification.

### ***Training Seminars***

SPSS Inc. 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, contact your local office, listed on the SPSS Web site at <http://www.spss.com/worldwide>.

### ***Technical Support***

The services of SPSS Technical Support are available to maintenance customers. Customers may contact Technical Support for assistance in using SPSS or for installation help for one of the supported hardware environments. To reach Technical Support, see the SPSS Web site at <http://www.spss.com>, or contact your local office, listed on the SPSS Web site at <http://www.spss.com/worldwide>. Be prepared to identify yourself, your organization, and the serial number of your system.

### ***Additional Publications***

Additional copies of SPSS product manuals may be purchased directly from SPSS Inc. Visit the SPSS Web Store at <http://www.spss.com/estore>, or contact your local SPSS office, listed on the SPSS Web site at <http://www.spss.com/worldwide>. For telephone orders in the United States and Canada, call SPSS Inc. at 800-543-2185. For telephone orders outside of North America, contact your local office, listed on the SPSS Web site.

The *SPSS Statistical Procedures Companion*, by Marija Norušis, has been published by Prentice Hall. A new version of this book, updated for SPSS 15.0, is planned. The *SPSS Advanced Statistical Procedures Companion*, also based on SPSS 15.0, is forthcoming. The *SPSS Guide to Data Analysis* for SPSS 15.0 is also in development. Announcements of publications available exclusively through Prentice Hall will be available on the SPSS Web site at <http://www.spss.com/estore> (select your home country, and then click Books).

### ***Tell Us Your Thoughts***

Your comments are important. Please let us know about your experiences with SPSS products. We especially like to hear about new and interesting applications using the SPSS Data Preparation add-on module. Please send e-mail to [suggest@spss.com](mailto:suggest@spss.com) or write to SPSS Inc., Attn.: Director of Product Planning, 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.

### ***About This Manual***

This manual documents the graphical user interface for the procedures included in the SPSS Data Preparation add-on module. Illustrations of dialog boxes are taken from SPSS for Windows. Dialog boxes in other operating systems are similar. Detailed information about the command syntax for features in the SPSS Data Preparation add-on module is available in two forms: integrated into the overall Help system and as a separate document in PDF form in the *SPSS 15.0 Command Syntax Reference*, available from the Help menu.

### ***Contacting SPSS***

If you would like to be on our mailing list, contact one of our offices, listed on our Web site at <http://www.spss.com/worldwide>.

---

# Contents

## **Part I: User's Guide**

<b>1</b>	<b>Introduction to SPSS Data Preparation</b>	<b>1</b>
	Usage of Data Preparation Procedures . . . . .	1
<b>2</b>	<b>Validation Rules</b>	<b>3</b>
	Load Predefined Validation Rules . . . . .	3
	Define Validation Rules . . . . .	4
	Define Single-Variable Rules . . . . .	5
	Define Cross-Variable Rules . . . . .	8
<b>3</b>	<b>Validate Data</b>	<b>10</b>
	Validate Data Basic Checks . . . . .	12
	Validate Data Single-Variable Rules . . . . .	14
	Validate Data Cross-Variable Rules . . . . .	15
	Validate Data Output . . . . .	16
	Validate Data Save . . . . .	18
<b>4</b>	<b>Identify Unusual Cases</b>	<b>20</b>
	Identify Unusual Cases Output . . . . .	23

Identify Unusual Cases Save . . . . .	25
Identify Unusual Cases Missing Values . . . . .	26
Identify Unusual Cases Options . . . . .	27
DETECTANOMALY Command Additional Features . . . . .	28

## **5 Optimal Binning 30**

Optimal Binning Output . . . . .	32
Optimal Binning Save . . . . .	33
Optimal Binning Missing Values . . . . .	34
Optimal Binning Options . . . . .	35
OPTIMAL BINNING Command Additional Features . . . . .	36

## **Part II: Examples**

### **6 Validate Data 38**

Validating a Medical Database . . . . .	38
Performing Basic Checks . . . . .	38
Copying and Using Rules from Another File . . . . .	42
Defining Your Own Rules . . . . .	54
Cross-Variable Rules . . . . .	61
Case Report . . . . .	62
Summary . . . . .	63
Related Procedures . . . . .	63

## **7 Identify Unusual Cases 64**

Identify Unusual Cases Algorithm . . . . .	64
Identifying Unusual Cases in a Medical Database . . . . .	65
Running the Analysis . . . . .	65
Case Processing Summary . . . . .	71
Anomaly Case Index List . . . . .	71
Anomaly Case Peer ID List . . . . .	72
Anomaly Case Reason List . . . . .	73
Scale Variable Norms . . . . .	75
Categorical Variable Norms . . . . .	76
Anomaly Index Summary . . . . .	78
Reason Summary . . . . .	79
Scatterplot of Anomaly Index by Variable Impact . . . . .	80
Summary . . . . .	83
Related Procedures . . . . .	83

## **8 Optimal Binning 84**

The Optimal Binning Algorithm . . . . .	84
Using Optimal Binning to Discretize Loan Applicant Data . . . . .	84
Running the Analysis . . . . .	85
Descriptive Statistics . . . . .	89
Model Entropy . . . . .	90
Binning Summaries . . . . .	91
Binned Variables . . . . .	96
Applying Syntax Binning Rules . . . . .	96
Summary . . . . .	98



***Index***

***100***



***Part I:***  
***User's Guide***



# ***Introduction to SPSS 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 SPSS Data Preparation add-on module allows you to identify unusual cases and invalid cases, variables, and data values in your active dataset.

## ***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.** 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. If your chosen predictive model requires categorical variables, discretize any scale variables.

Once your data file is “clean,” you are ready to build models from other SPSS 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.

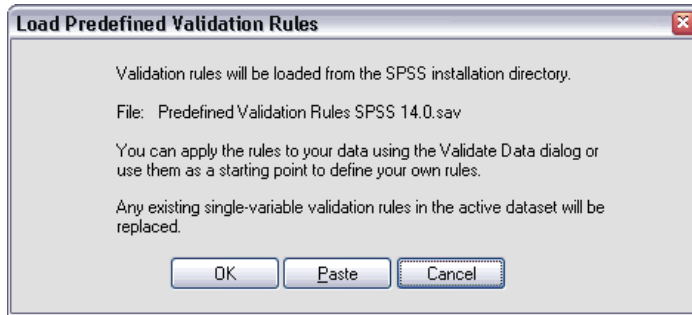
## ***Load Predefined Validation Rules***

You can quickly obtain a set of ready-to-use validation rules by loading predefined rules from an external data file that ships with SPSS.

### ***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...

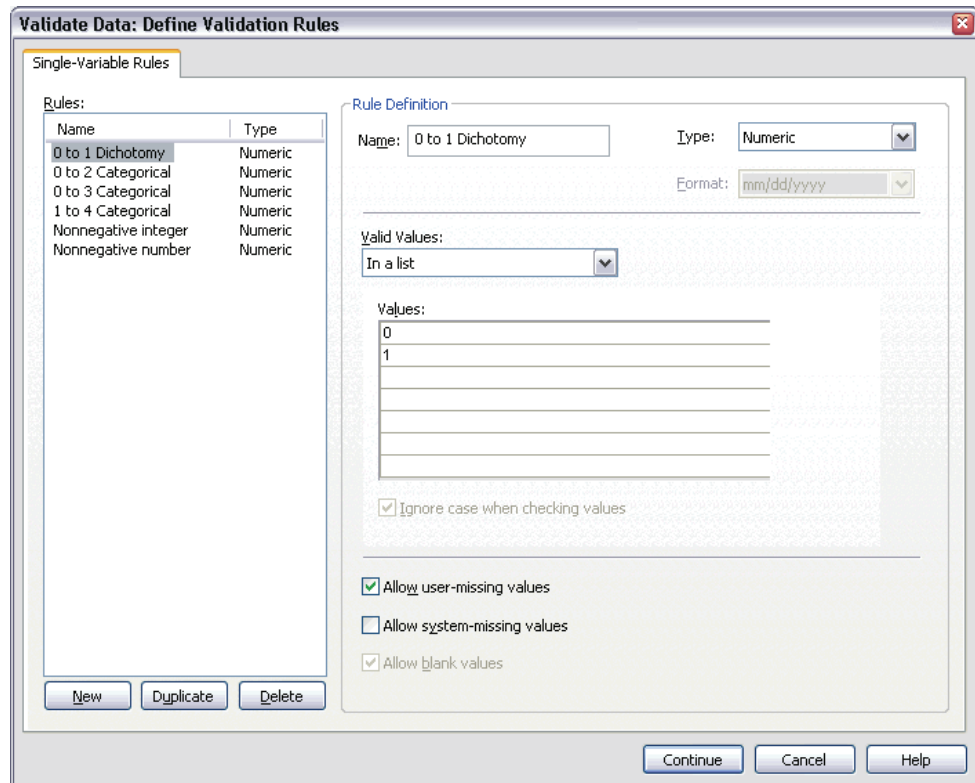
The dialog box is populated with single-variable and cross-variable validation rules read from the SPSS data dictionary. When there are no rules, a new placeholder rule that you can modify to suit your purposes is created automatically.

- ▶ 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 SPSS 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*

Valid Values: Within a range

Minimum: 0

Maximum:

Allow unlabeled values within range  
Since long string variables do not have value labels, you should always check this option for such variables.

Allow noninteger values within range

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*

The screenshot shows a dialog box titled 'Valid Values:'. At the top, there is a dropdown menu with the text 'In a list' and a downward arrow. Below this is a section labeled 'Values:' containing a table with five rows. The first row contains the value '0', and the second row contains the value '1'. The remaining three rows are empty. At the bottom of the dialog, there is a checked checkbox followed by the text 'Ignore case when checking values'.

Values:
0
1

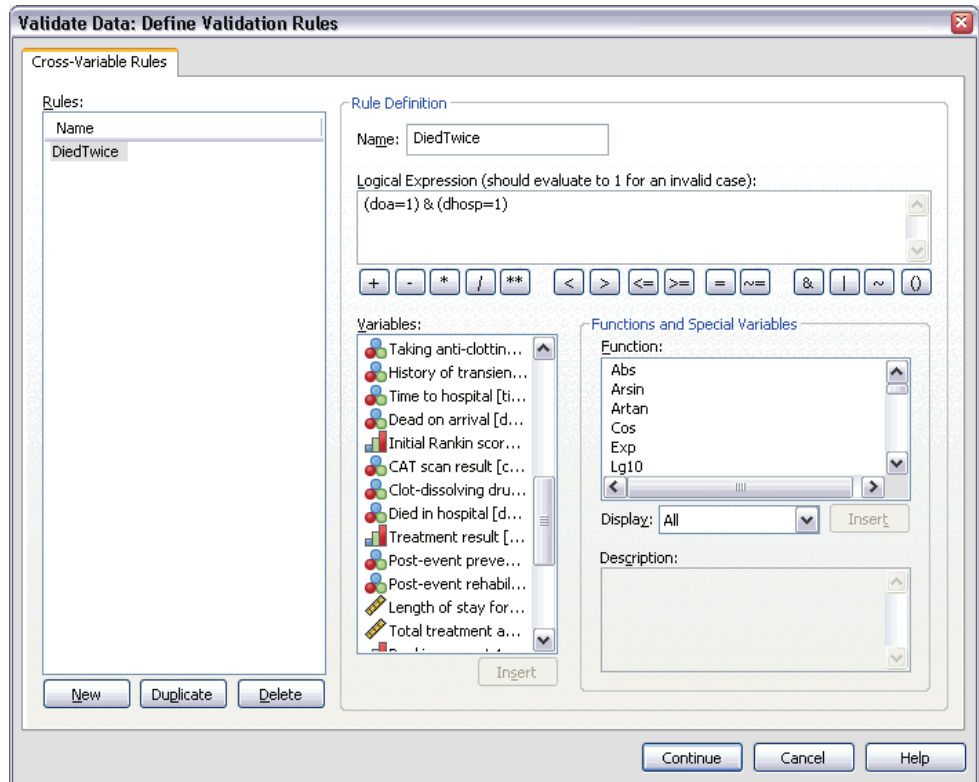
Ignore case when checking values

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.

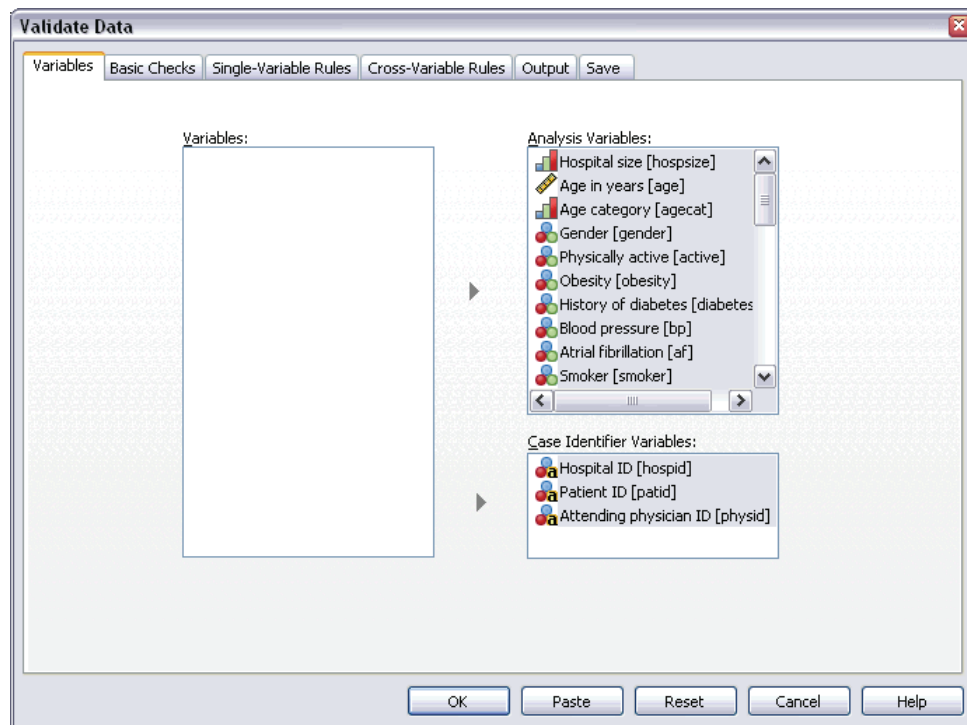
**Statistics.** The procedure produces lists of variables, cases, and data values that fail various checks, counts of violations of single-variable and cross-variable rules, and simple descriptive summaries of analysis variables.

**Weights.** The procedure ignores the SPSS 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:

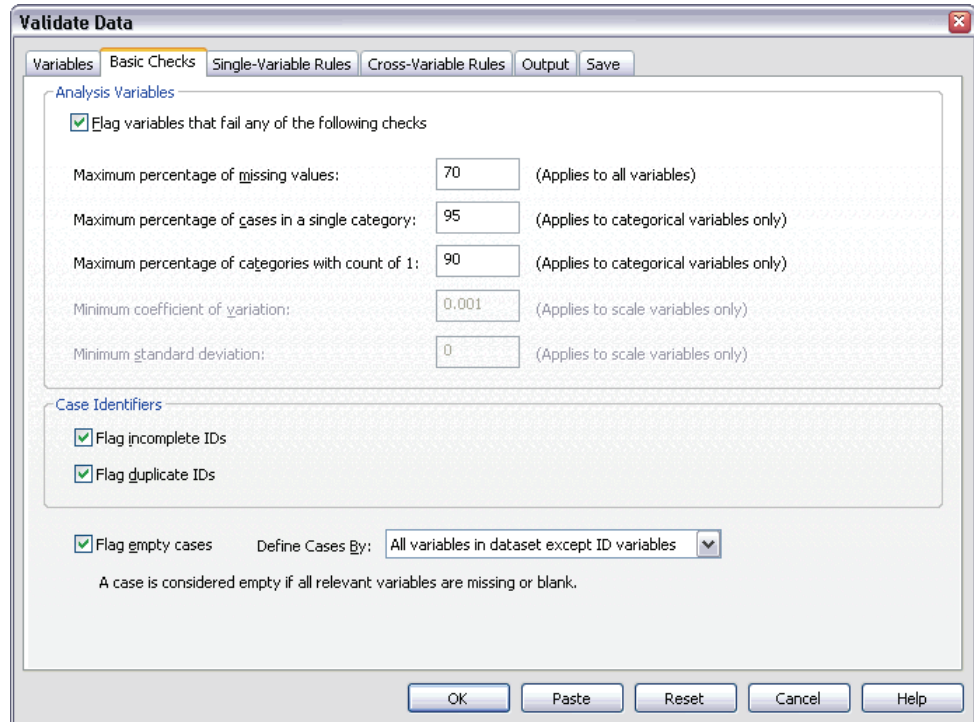
- ▶ Click the Cross-Variable Rules tab and apply one or more cross-variable rules.

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.

## Validate Data Basic Checks

Figure 3-2  
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.

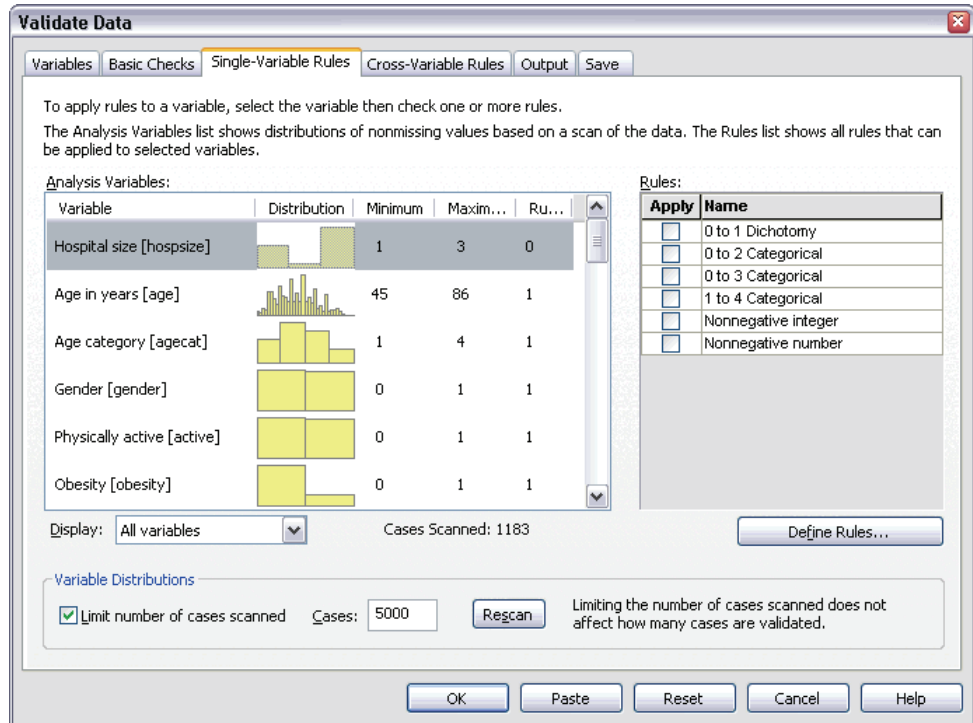
**Case Identifiers.** If you selected any case identifier variables on the Variables tab, you can select any of the following checks of their validity.

- **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 3-3  
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 Define Single-Variable Rules in Chapter 2 on p. 5.](#)

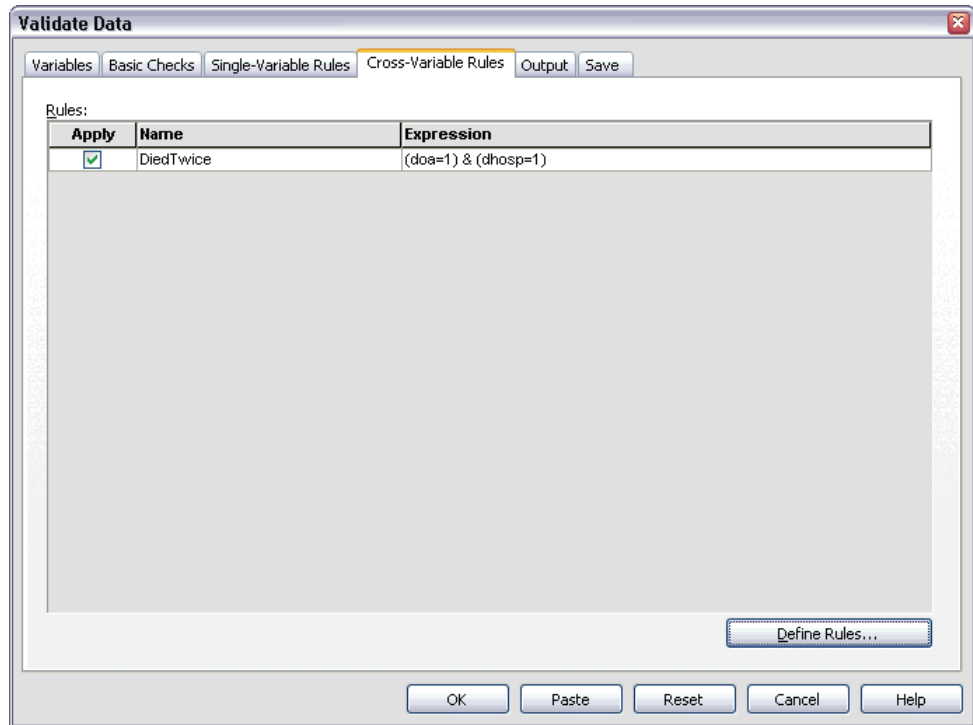
**Analysis Variables.** The list shows analysis variables, summarizes their distributions, and shows the number of rules applied to each variable. Note that user- and system-missing values are not included in the summaries. The Display drop-down list controls which variables are shown; you can choose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

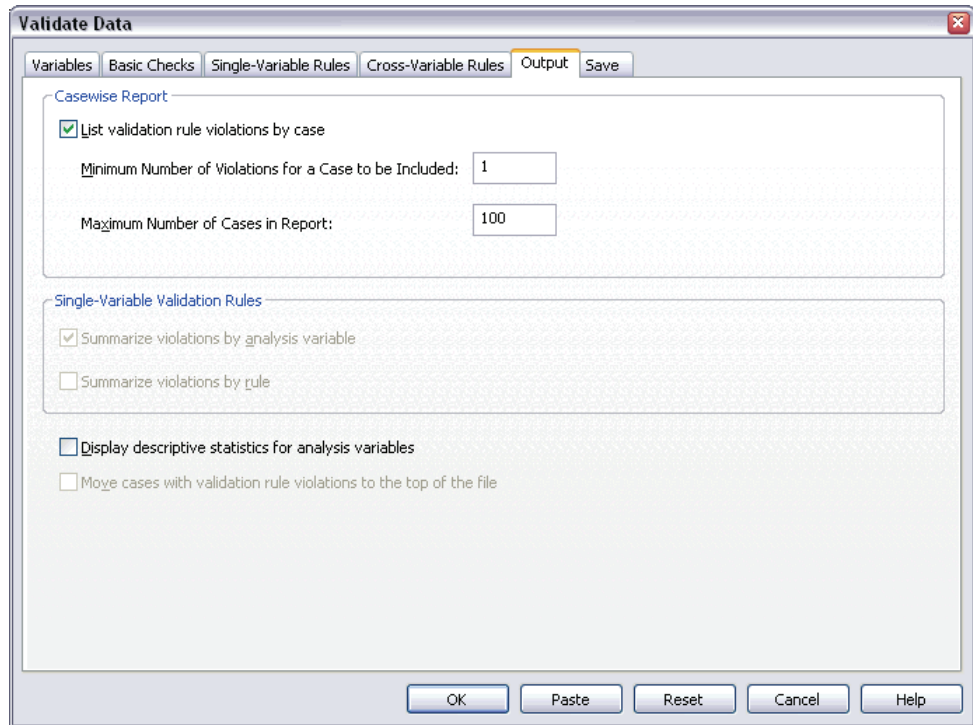
Figure 3-4  
Validate Data dialog box, Cross-Variable Rules tab



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 [Define Cross-Variable Rules in Chapter 2 on p. 8](#).

## Validate Data Output

Figure 3-5  
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.

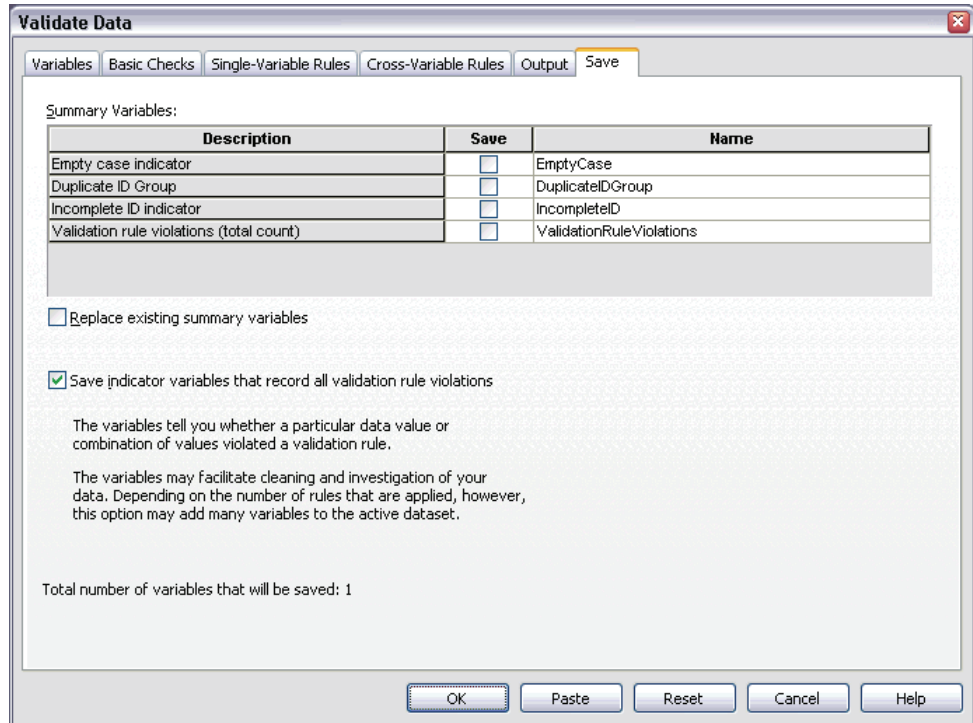
- **Summarize violations by analysis variable.** For each analysis variable, this option shows all single-variable validation rules that were violated and the number of values that violated each rule. It also reports the total number of single-variable rule violations for each variable.
- **Summarize 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 3-6  
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.

# ***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 SPSS 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.

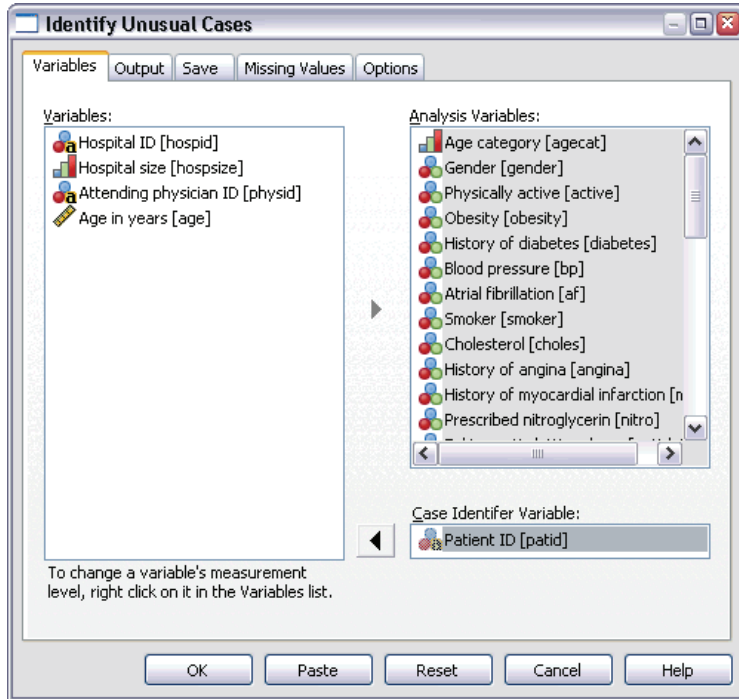
**Case order.** Note that the solution may depend on the order of cases. To minimize order effects, randomly order the cases. To verify the stability of a given solution, you may want to obtain several different solutions with cases sorted in different random orders. In situations with extremely large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

### ***To Identify Unusual Cases***

- ▶ From the menus choose:
  - Data
  - Identify Unusual Cases...

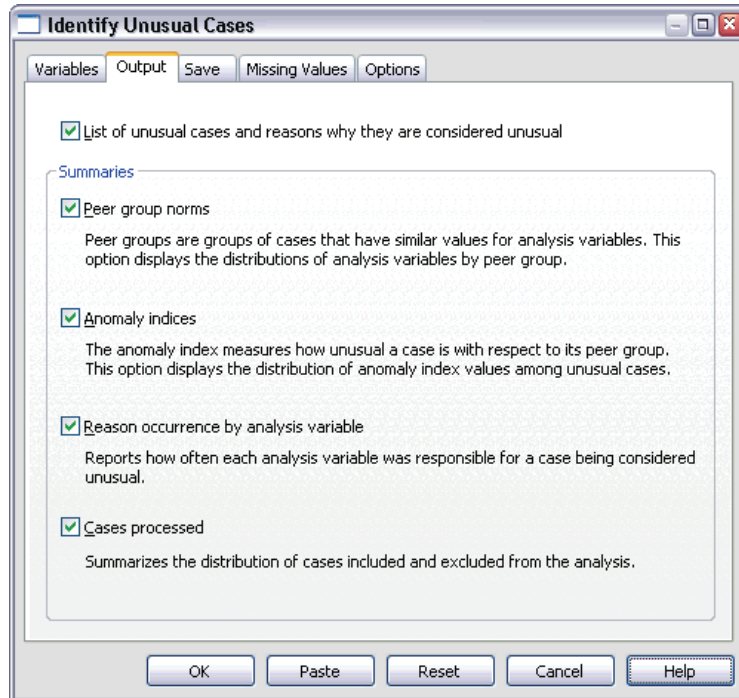
Figure 4-1  
*Identify Unusual Cases dialog box, Variables tab*



- ▶ Select at least one analysis variable.
- ▶ Optionally, choose a case identifier variable to use in labeling output.

## Identify Unusual Cases Output

Figure 4-2  
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.

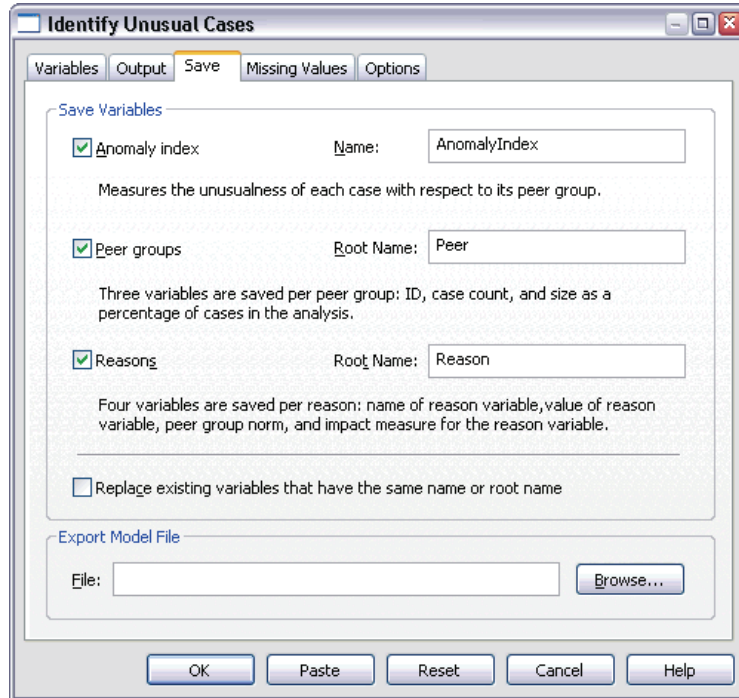
All tables are sorted by anomaly index in descending order. Moreover, the IDs of the cases are displayed if the case identifier variable is specified on the Variables tab.

**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 4-3  
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

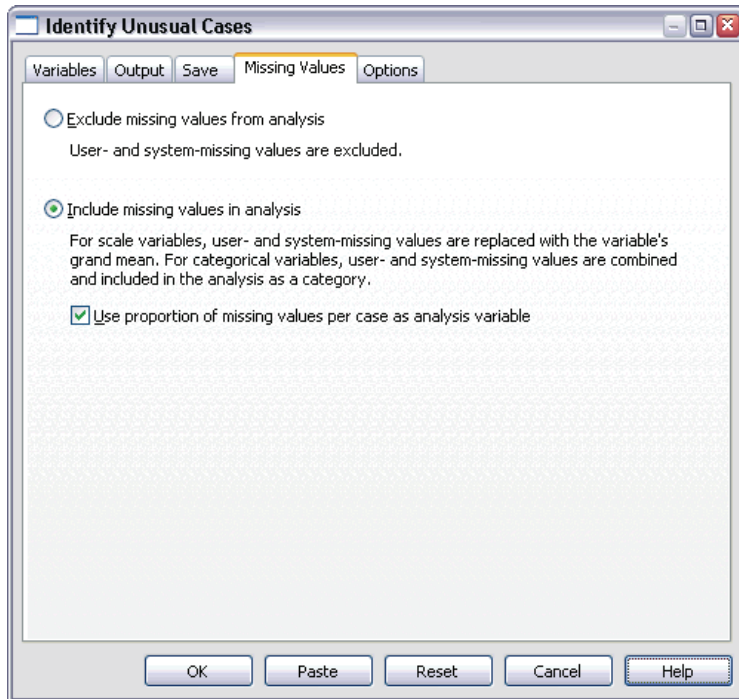
the peer group ID of the case, *PeerSize* is the group's size, and *PeerPctSize* is the group's size as a percentage.

- **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 4-4  
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 4-5  
Identify Unusual Cases dialog box, Options tab

The screenshot shows the 'Identify Unusual Cases' dialog box with the 'Options' tab selected. The dialog has five tabs: 'Variables', 'Output', 'Save', 'Missing Values', and 'Options'. The 'Options' tab contains the following settings:

- Criteria for Identifying Unusual Cases:**
  - Percentage of cases with highest anomaly index values
    - Percentage: 2
  - Fixed number of cases with highest anomaly index values
    - Number: [ ]
  - Identify only cases whose anomaly index value meets or exceeds a minimum value
    - Cutoff: 2
- Number of Peer Groups:**
  - Minimum: 1
  - Maximum: 15
- Maximum Number of Reasons:** 3

Specify the number of reasons reported in output and added to the active dataset if reason variables are saved. The value is adjusted downward if it exceeds the number of analysis variables.

Buttons at the bottom: OK, Paste, Reset, Cancel, Help.

**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 SPSS command 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 *SPSS 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 cut points 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 SPSS syntax for use in discretizing new data.

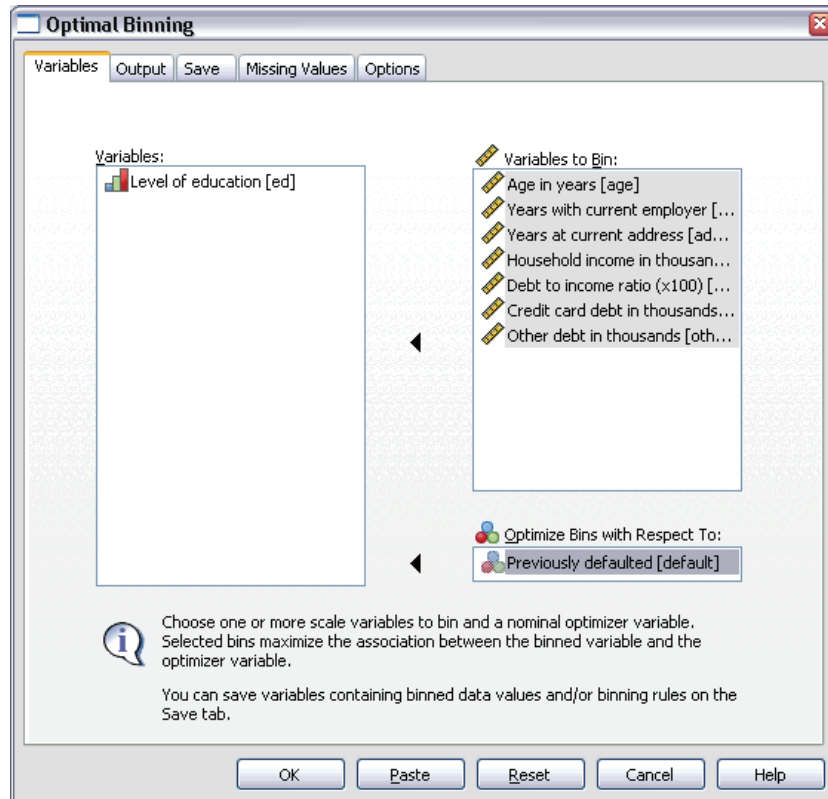
**Data.** This procedure expects the binning input variables to be scale, numeric variables. The guide variable should be categorical and can be string or numeric.

### To Obtain Optimal Binning

From the menus choose:

Transform  
Optimal Binning...

Figure 5-1  
Optimal Binning dialog box, Variables tab

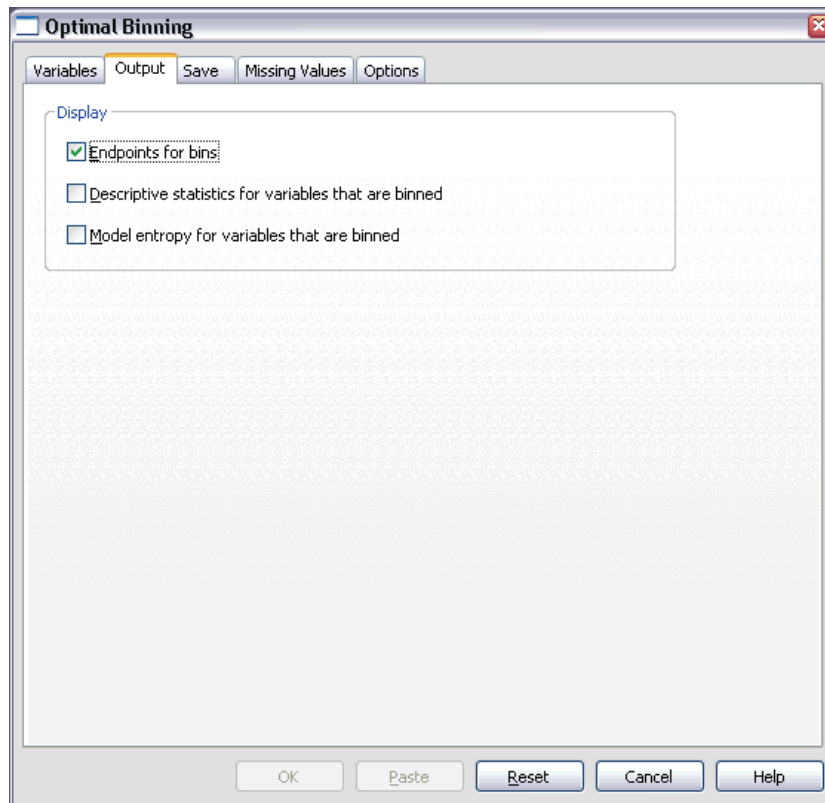


- ▶ Select one or more binning input variables.
- ▶ Select a guide variable.

Variables containing the binned data values are not generated by default. Use the [Save](#) tab to save these variables.

## Optimal Binning Output

Figure 5-2  
Optimal Binning dialog box, Output tab



The Output tab controls the display of the results.

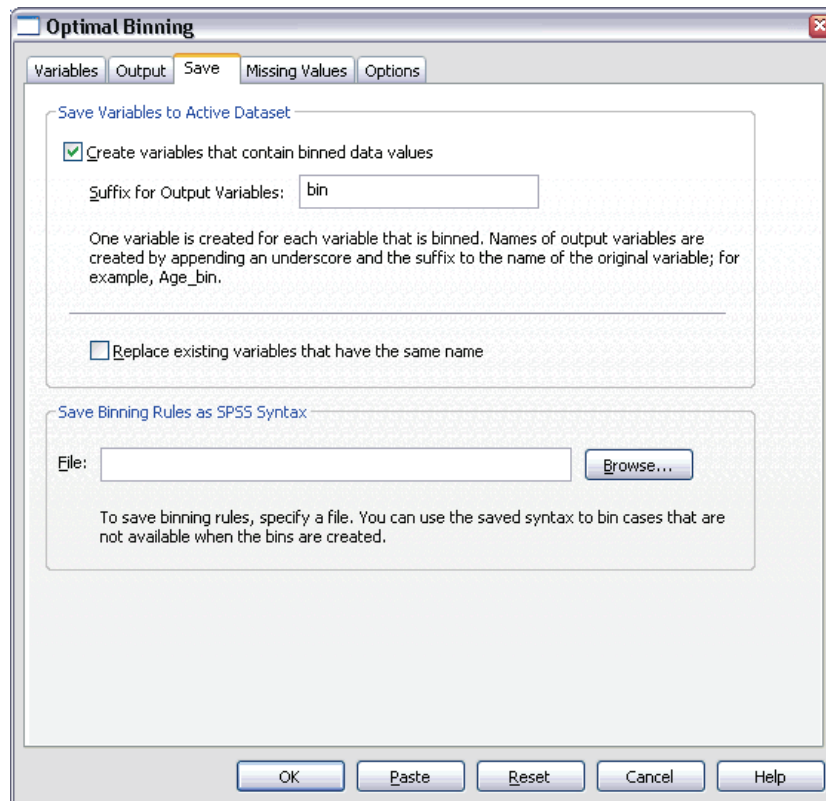
- **Endpoints for bins.** Displays the set of endpoints for each binning input variable.
- **Descriptive statistics for variables that are binned.** For each binning input variable, this option displays the number of cases with valid values, the number of cases with missing values, the number of distinct 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 5-3  
Optimal Binning dialog box, Save tab

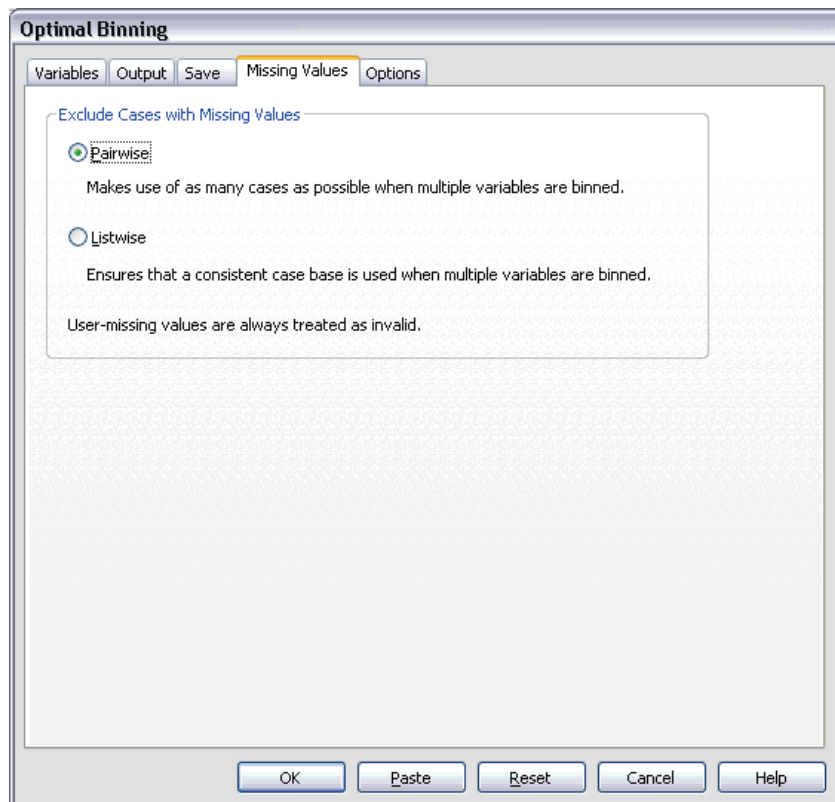


**Save Variables to Active Dataset.** Variables containing the binned data values can be used in place of the original variables in further analysis.

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

## ***Optimal Binning Missing Values***

Figure 5-4  
*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 5-5  
Optimal Binning dialog box, Options tab

**Optimal Binning**

Variables Output Save Missing Values Options

**Preprocessing**

Pre-bin variables to improve performance with large datasets  
Specify the maximum number of bins that any variable should end up with after preprocessing.

Maximum Number of Bins:

**Sparsely Populated Bins**

Merge bins that have relatively small case counts with a larger neighbor

Threshold (ratio):

A bin is merged if the ratio of its size (number of cases) to that of a neighboring bin is smaller than the specified threshold. Larger thresholds tend to result in more merging.

**Bin Endpoints**

Lower endpoint is inclusive, upper is exclusive (lower  $\leq$  x < upper)

Lower endpoint is exclusive, upper is inclusive (lower < x  $\leq$  upper)

**First (Lowest) Bin**

Unbounded (extends to negative infinity)

Bounded by lowest data value

**Last (Highest) Bin**

Unbounded (extends to positive infinity)

Bounded by highest data value

OK Paste Reset Cancel Help

**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 cut points:

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

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

where  $\text{sizeof}(b)$  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 cut points, this is largely a matter of preference.

**First (Lowest) / Last (Highest) Bin.** These options specify how the minimum and maximum cut points 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 SPSS command language also allows you to:

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

See the *SPSS 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***

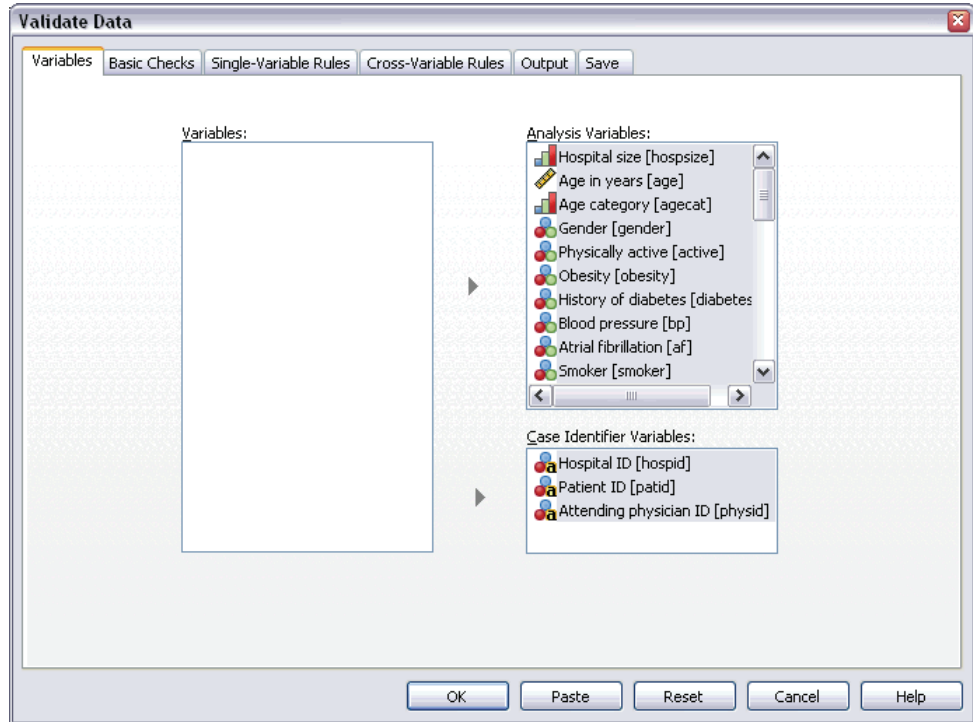
An analyst hired by a medical group must maintain the quality of the information in the system. This process involves checking the values and variables and preparing a report for the manager of the data entry team.

The latest state of the database is collected in *stroke\_invalid.sav*. 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***

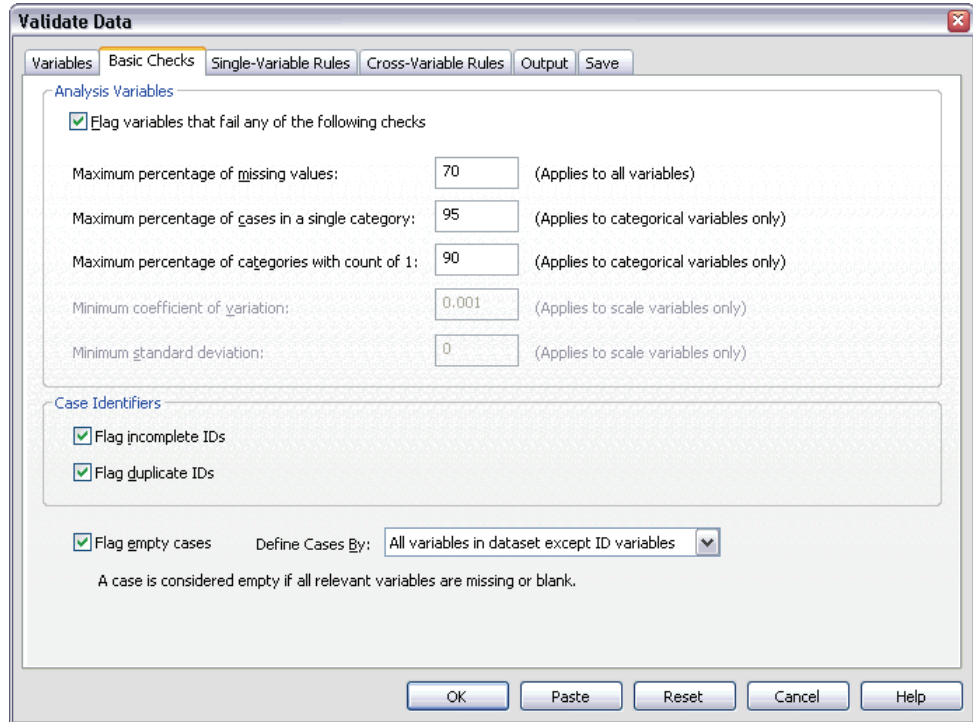
- ▶ To run a Validate Data analysis, from the menus choose:
  - Data
  - Validation
  - Validate Data...

Figure 6-1  
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 6-2  
Validate Data dialog box, Basic Checks tab



The default settings are the settings you want to run.

- ▶ Click OK.

## Warnings

Figure 6-3  
Warnings

Some or all requested output is not displayed because all cases, variables, or data values passed the requested checks.

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 6-4  
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 6-5  
Duplicate case identifiers (first 11 shown)

Duplicate Identifiers Group	Number of Duplicates	Cases with Duplicate Identifiers	Identifier		
			hospid	patid	physid
1	2	10, 11	PEW	1406462419	355184
2	2	14, 15	PEW	2191527525	355184
3	2	21, 22	PEW	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

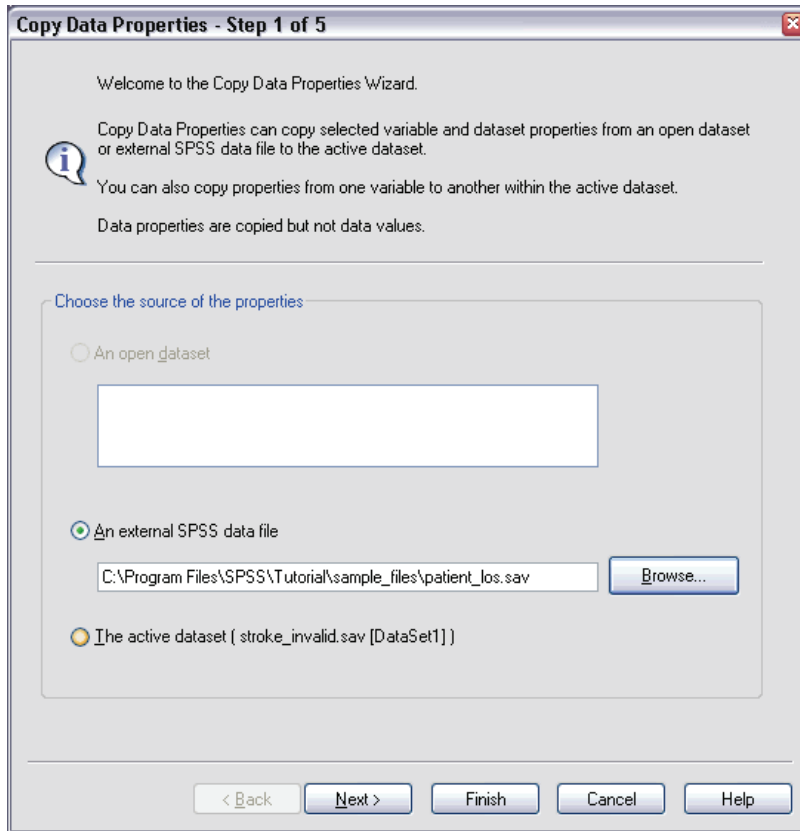
A case should be uniquely identified by the combination of values of the identifier variables. The first 11 entries in the duplicate identifiers table are shown here. These duplicates are patients with multiple events who were entered as separate cases for each event. Because this information can be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

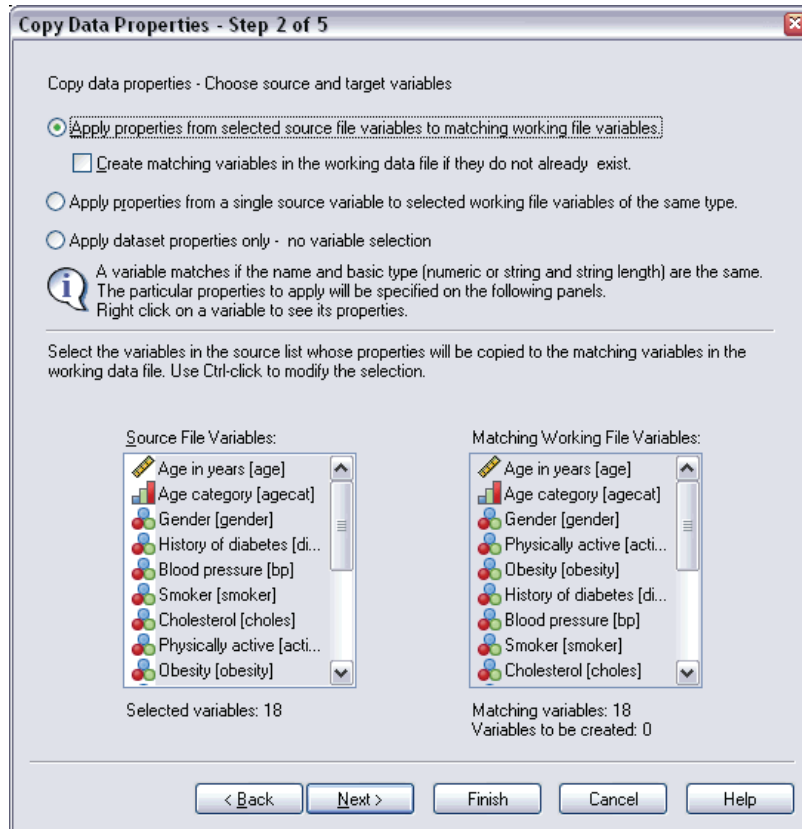
- ▶ To copy rules from another file, from the menus choose:
  - Data
  - Copy Data Properties...

Figure 6-6  
Copy Data Properties, Step 1 (welcome)



- ▶ Choose to copy properties from an external SPSS data file, *patient\_los.sav*, which can be found in the *\Tutorial\sample\_files* subdirectory of the SPSS installation directory.
- ▶ Click Next.

Figure 6-7  
Copy Data Properties, Step 2 (choose variables)

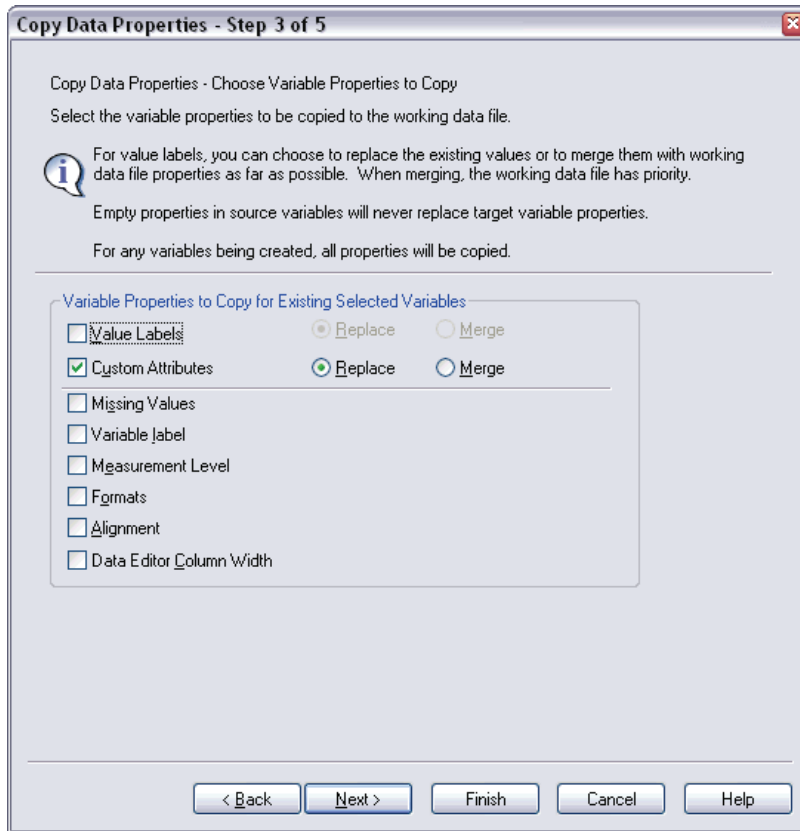


These are the variables whose properties you want to copy from *patient\_los.sav* to the corresponding variables in *stroke\_invalid.sav*.

- ▶ Click Next.

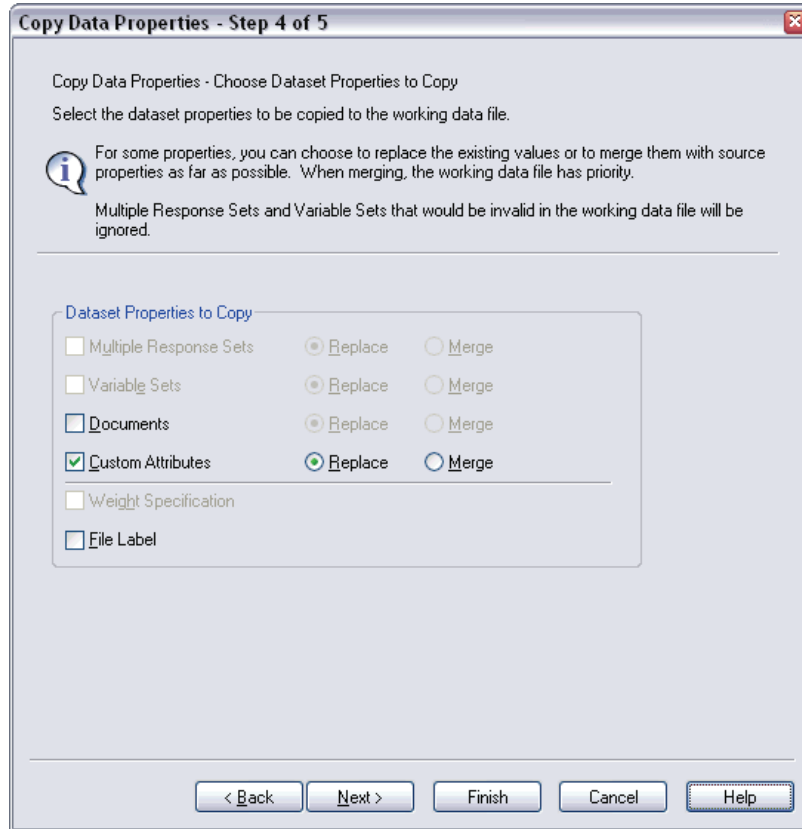


**Figure 6-8**  
*Copy Data Properties, Step 4 (choose variable properties)*



- ▶ Deselect all properties except Custom Attributes.
- ▶ Click Next.

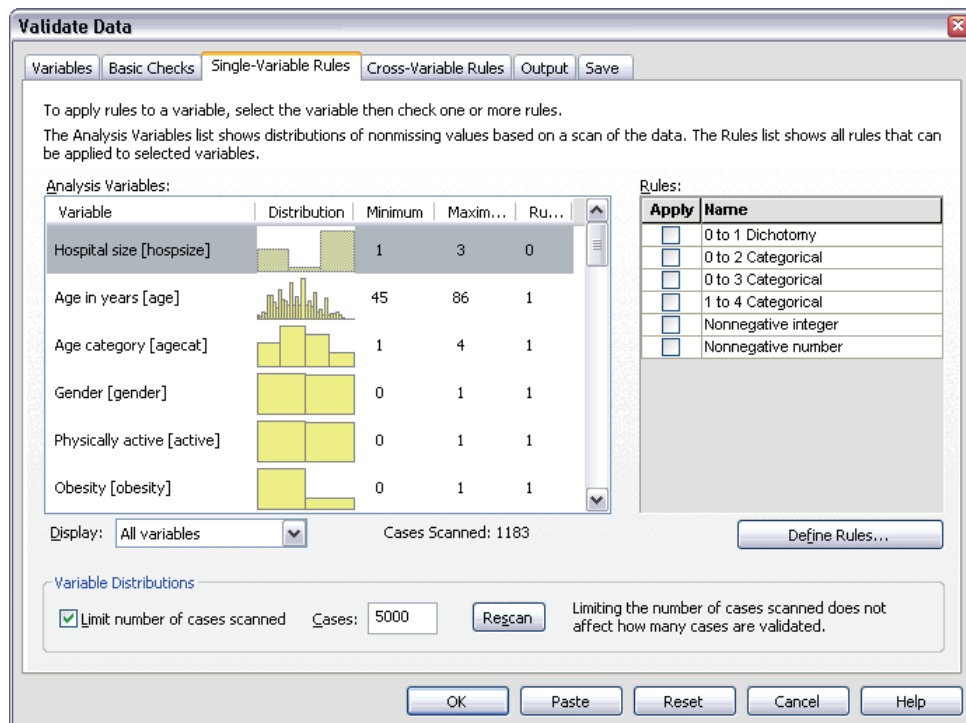
**Figure 6-9**  
*Copy Data Properties, Step 4 (choose dataset properties)*



- ▶ Select Custom Attributes.
- ▶ Click Finish.

You are now ready to reuse the validation rules.

Figure 6-10  
Validate Data dialog box, Single-Variable Rules tab

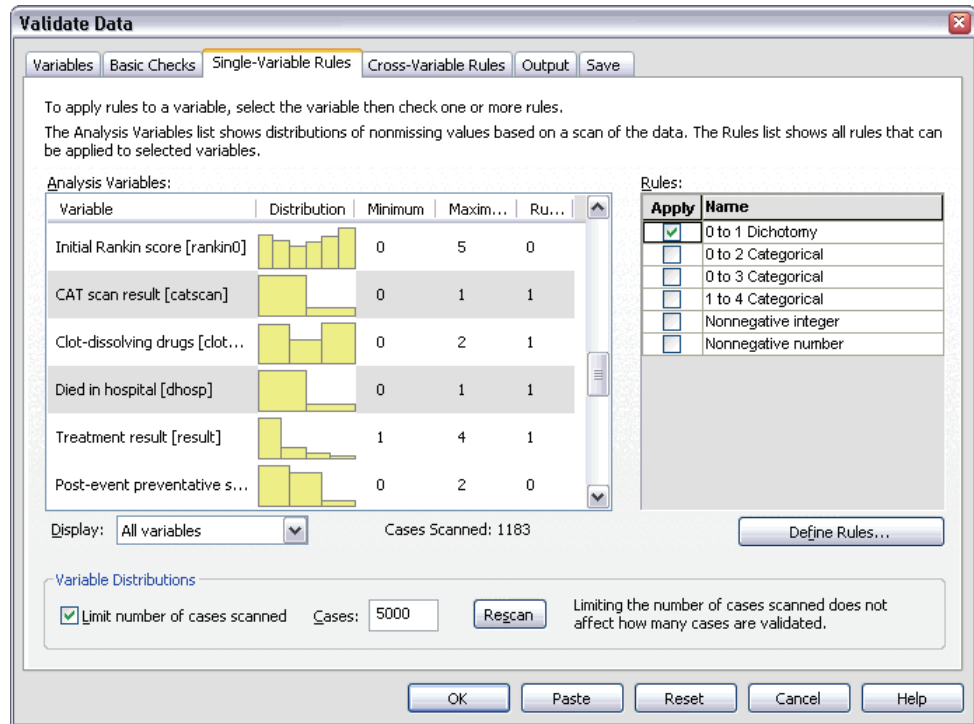


- ▶ To validate the *stroke\_invalid.sav* data by using the copied rules, click the Dialog Recall toolbar button and choose Validate Data.
- ▶ Click the Single-Variable Rules tab.

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.

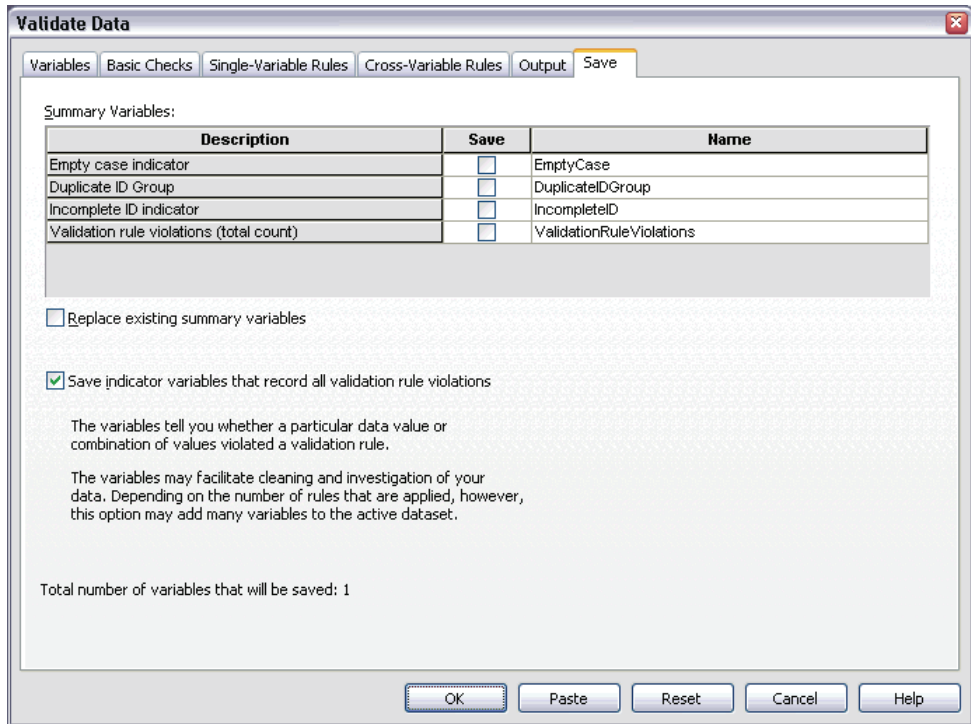
The Rules list shows the single-variable validation rules that are available in the data file. These rules were all copied from *patient\_los.sav*. Note that some of these rules are applicable to variables that did not have exact counterparts in the other data file.

Figure 6-11  
 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 6-12  
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 6-13  
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 6-14  
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 6-15  
Case report

Case	Validation Rule	Identifier		
	Single-Variable <sup>a</sup>	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

<sup>a</sup>. 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 6-16  
Data Editor with saved indicators of rule violations

	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia	@0to1Dichotomy_tom
1	4	.00	.00	.00	.00	.00	.00
2	4	.00	.00	.00	.00	.00	.00
3	1	.00	.00	.00	.00	.00	.00
4	4	.00	.00	.00	.00	.00	.00
5	3	.00	.00	.00	.00	.00	.00
6	4	.00	.00	.00	.00	.00	.00
7	4	.00	.00	.00	.00	.00	.00
8	4	.00	.00	.00	.00	.00	.00
9	4	.00	.00	.00	.00	.00	.00
10	2	.00	.00	.00	.00	.00	.00
11	2	nn	nn	nn	nn	nn	nn

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 6-17  
Data Editor with indicator of rule violation for case 175

The screenshot shows the SPSS Data Editor window for the file 'stroke\_invalid.sav'. The window title is '\*stroke\_invalid.sav [] - SPSS Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. The toolbar contains various icons for file operations and data manipulation. The main area displays a data table with the following columns: recbar3, @Oto1Dichotomy\_doa, @Oto1Dichotomy\_gender, @Oto1Dichotomy\_angina, @1to4Categorical\_agecat, and Nonnegativeinteger\_time. The row for case 175 is highlighted, and the value for the variable @Oto1Dichotomy\_angina is 1.00, which is a rule violation. The status bar at the bottom indicates 'SPSS Processor is ready'.

	recbar3	@Oto1Dichotomy_doa	@Oto1Dichotomy_gender	@Oto1Dichotomy_angina	@1to4Categorical_agecat	Nonnegativeinteger_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

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 6-18  
Data Editor with invalid value for History of angina

The screenshot shows the SPSS Data Editor window for a file named \*stroke\_invalid.sav. The window title is "\*stroke\_invalid.sav [] - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. The toolbar contains various icons for file operations and data manipulation. The main area displays a data table with the following columns: af, smoker, choles, angina, mi, nitro, anticlot, and tia. The row labels on the left are 172 through 180. The value -1 is entered in the 'angina' column for row 175. The status bar at the bottom indicates "SPSS Processor is ready".

	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

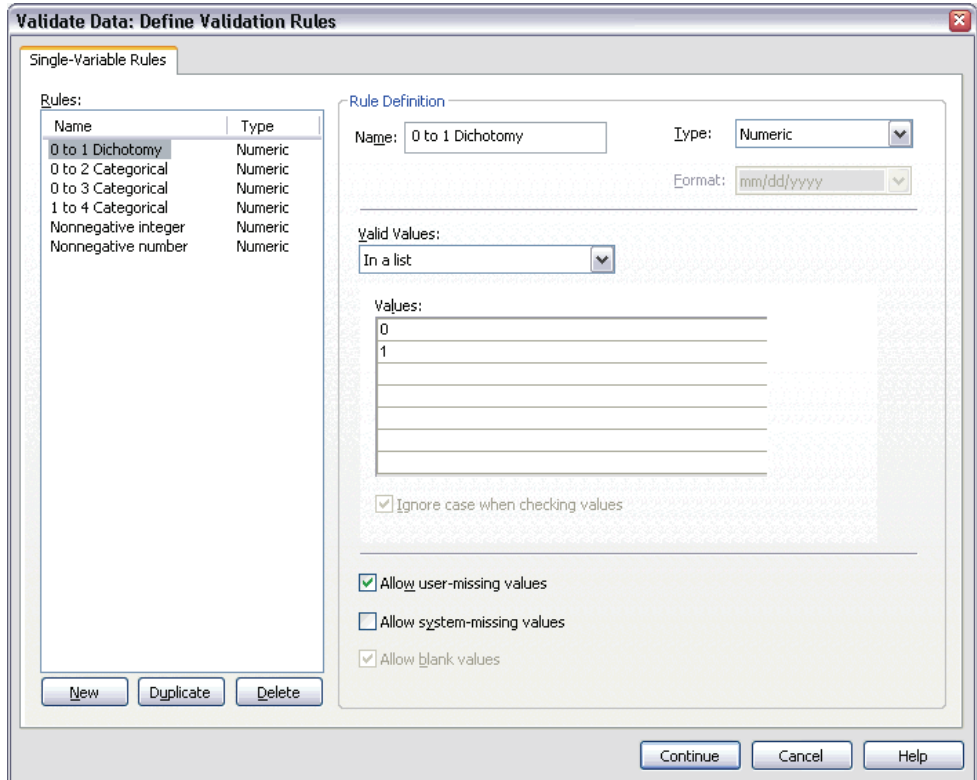
History of angina has a value of -1. While this value is a valid missing value for treatment and result variables in the data file, it is invalid here because the patient history values do not currently have user-missing values defined.

## 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.

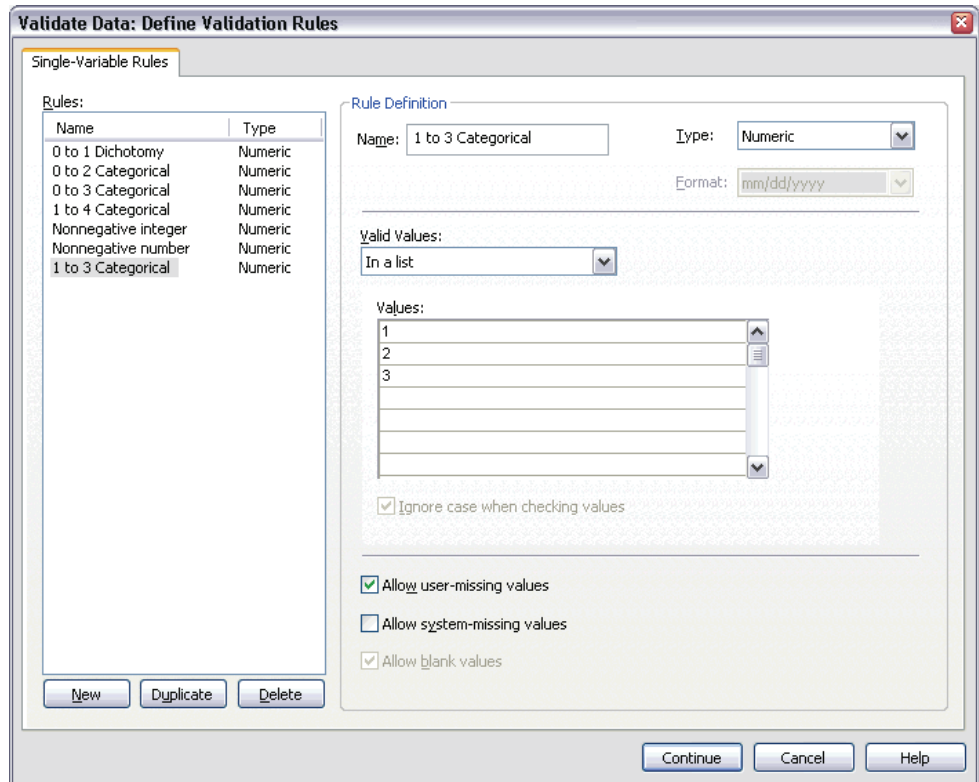
**Figure 6-19**  
*Define Validation Rules dialog box, Single-Variable Rules tab*



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.

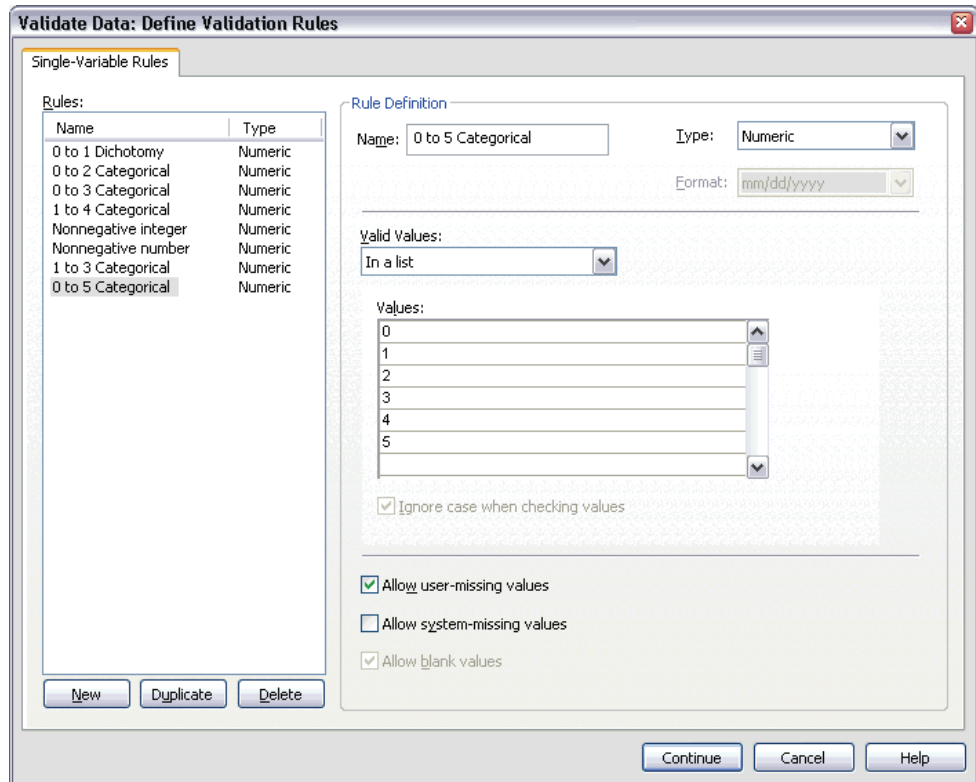
- To define a rule, click New.

**Figure 6-20**  
*Define Validation Rules dialog box, Single-Variable Rules tab (1 to 3 Categorical defined)*



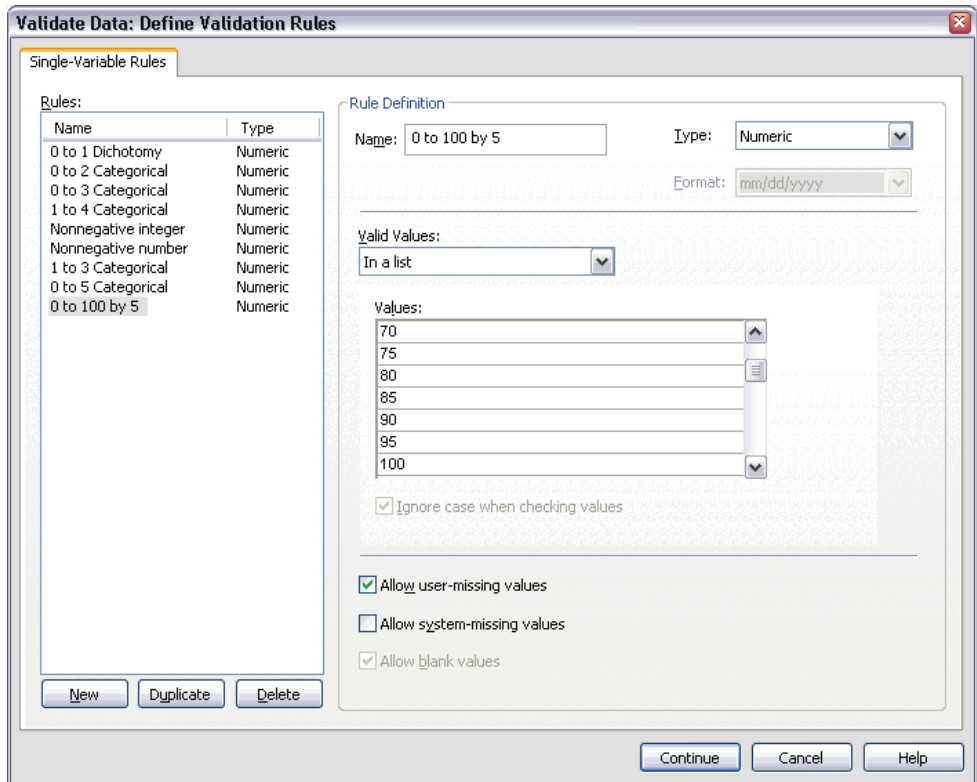
- ▶ Type 1 to 3 Categorical as the rule name.
- ▶ For Valid Values, choose In a list.
- ▶ Type 1, 2, and 3 as the values.
- ▶ Deselect Allow system-missing values.
- ▶ To define the rule for Rankin scores, click New.

**Figure 6-21**  
*Define Validation Rules dialog box, Single-Variable Rules tab (0 to 5 Categorical defined)*



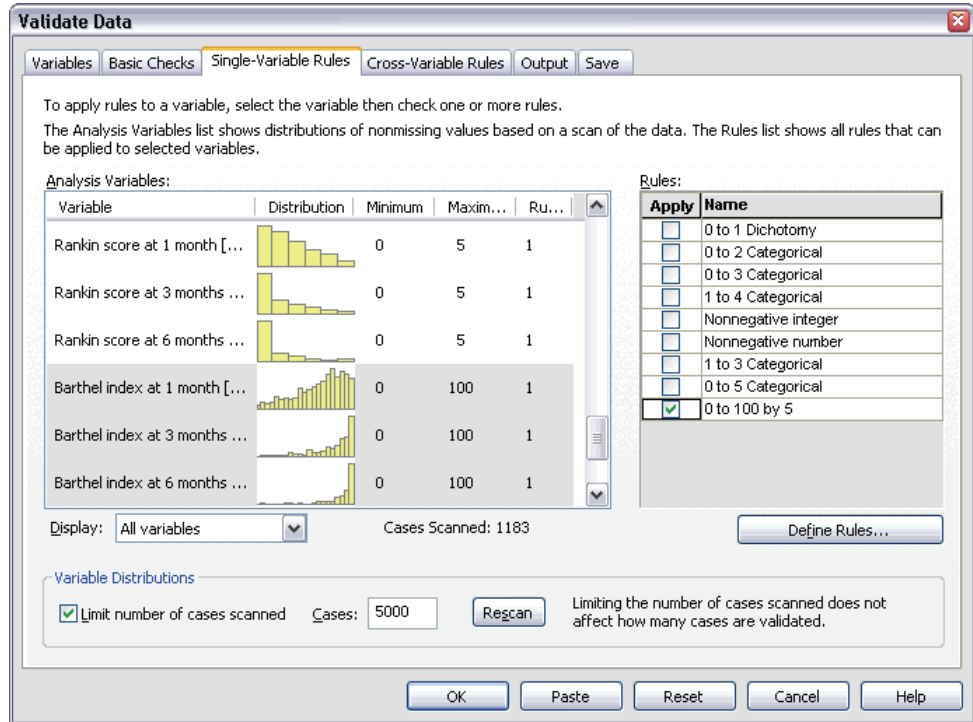
- ▶ Type 0 to 5 Categorical as the rule name.
- ▶ For Valid Values, choose In a list.
- ▶ Type 0, 1, 2, 3, 4, and 5 as the values.
- ▶ Deselect Allow system-missing values.
- ▶ To define the rule for Barthel indices, click New.

**Figure 6-22**  
*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 6-23  
Validate Data dialog box, Single-Variable Rules tab (0 to 100 by 5 defined)



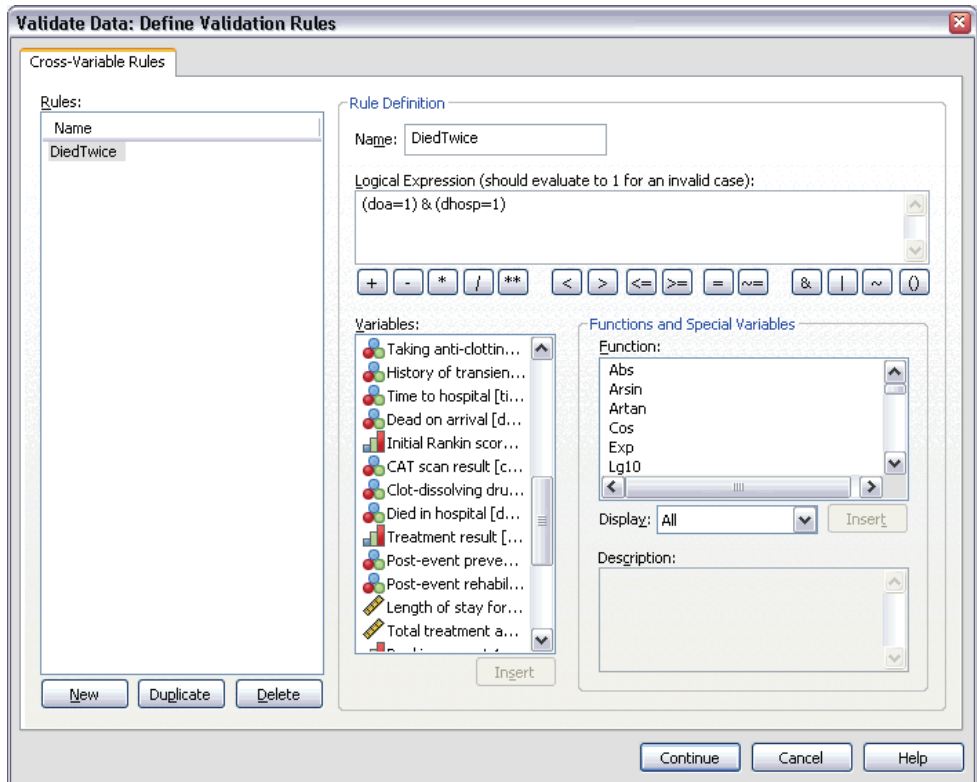
Now you need to apply the defined rules to analysis variables.

- ▶ 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 6-24**  
*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 6-25  
*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 6-26  
Case report

Case	Validation Rule Violations		Identifier		
	Single-Variable <sup>a</sup>	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	VWL	9714672452	237547
881		Died twice	VWL	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.

# ***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***

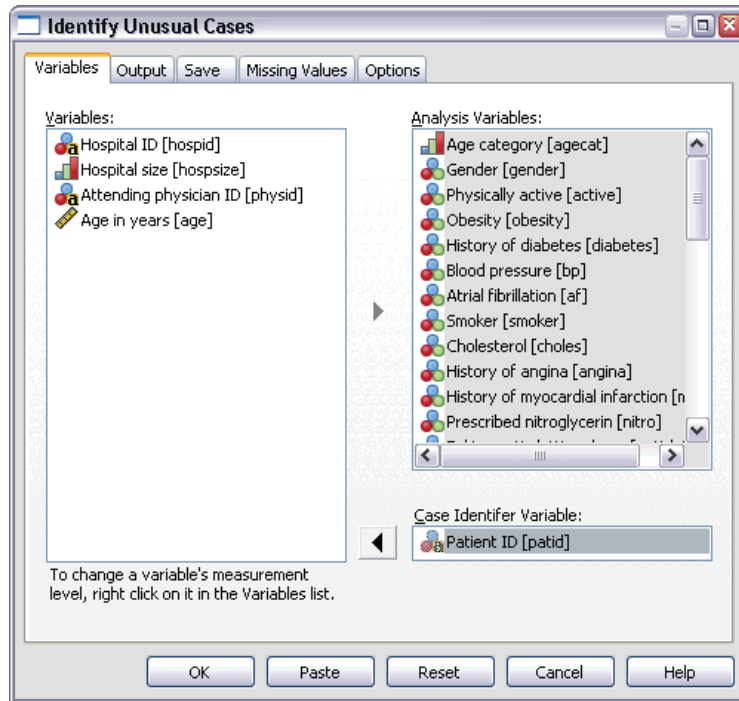
A data analyst hired to build predictive models for stroke treatment outcomes is concerned about data quality because such models can be sensitive to unusual observations. Some of these outlying observations represent truly unique cases and are thus unsuitable for prediction, while other observations 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*. 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***

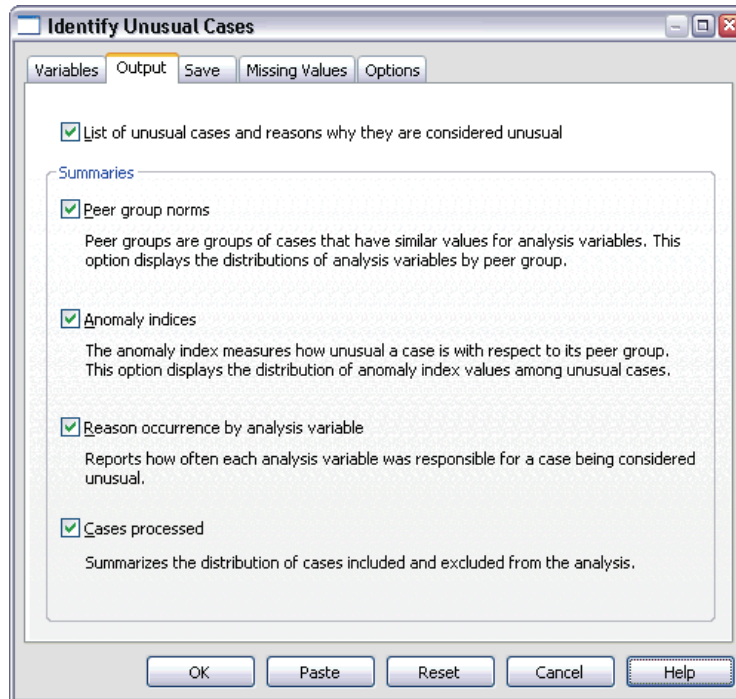
- ▶ To identify unusual cases, from the menus choose:
  - Data
  - Identify Unusual Cases...

Figure 7-1  
*Identify Unusual Cases dialog box, Variables tab*



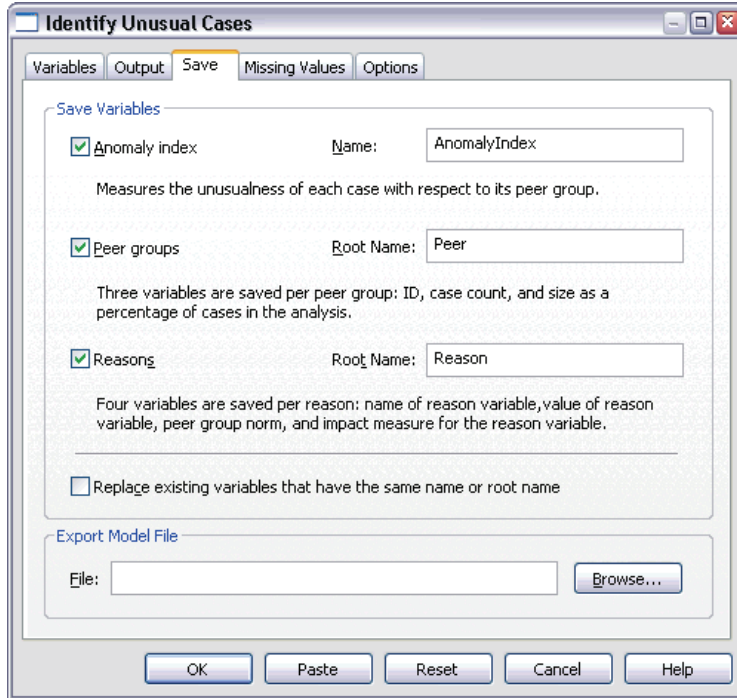
- ▶ Select *Age category* through *Stroke between 3 and 6 months* as analysis variables.
- ▶ Select *Patient ID* as the case identifier variable.
- ▶ Click the *Output* tab.

Figure 7-2  
*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 7-3  
*Identify Unusual Cases dialog box, Save tab*



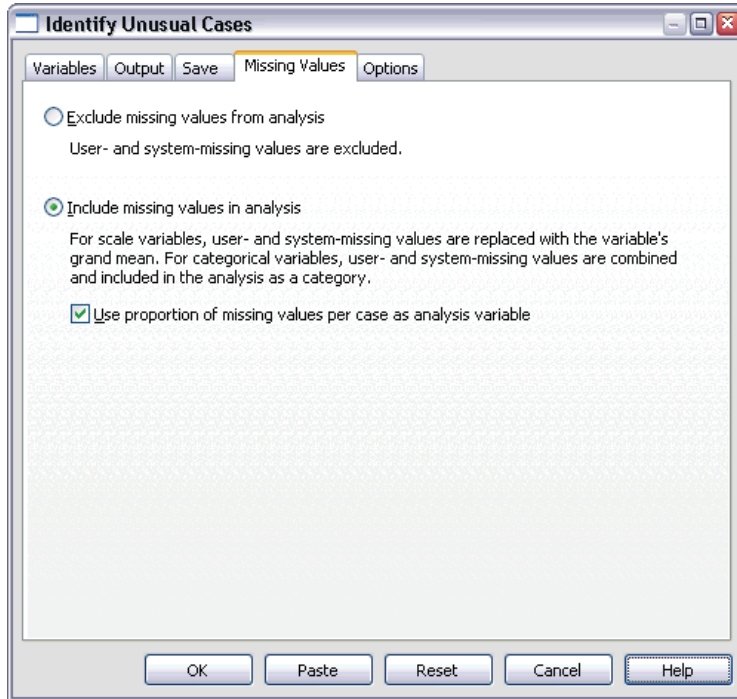
- ▶ Select Anomaly index, Peer groups, and Reasons.

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

- ▶ Click the Missing Values tab.

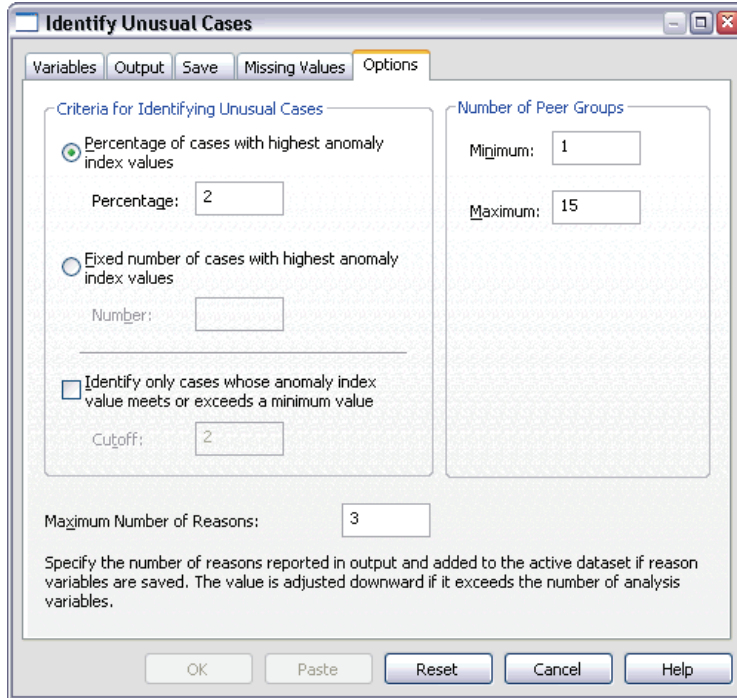


Figure 7-4  
*Identify Unusual Cases dialog box, Missing Values tab*



- ▶ 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 7-5  
*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 7-6  
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 7-7  
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 7-8  
*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	0816889249	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 7-9  
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	0816889249	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 7-10  
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	rechart1	.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 7-11  
Scale variable norms

		Peer ID			Combined
		1	2	3	
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 7-12  
Categorical variable norms (first 10 variables)

		Peer ID			Combined
		1	2	3	
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 7-13  
Categorical variable norms (selected variables)

		Peer ID			Combined
		1	2	3	
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 7-14  
*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 7-15  
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

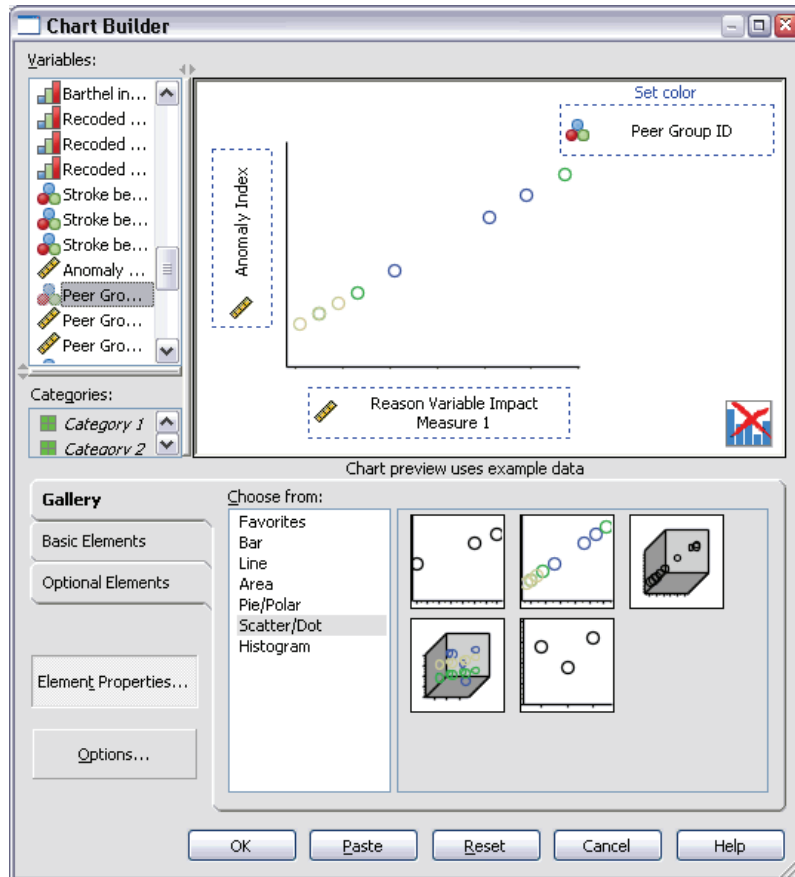
For each variable in the analysis, the table summarizes the variable's role as a primary reason. Most variables, such as variables from *Dead on arrival* to *Post-event rehabilitation*, are not the primary reason that any of the cases are on the anomaly list. *Barthel index at 1 month* is the most frequent 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...

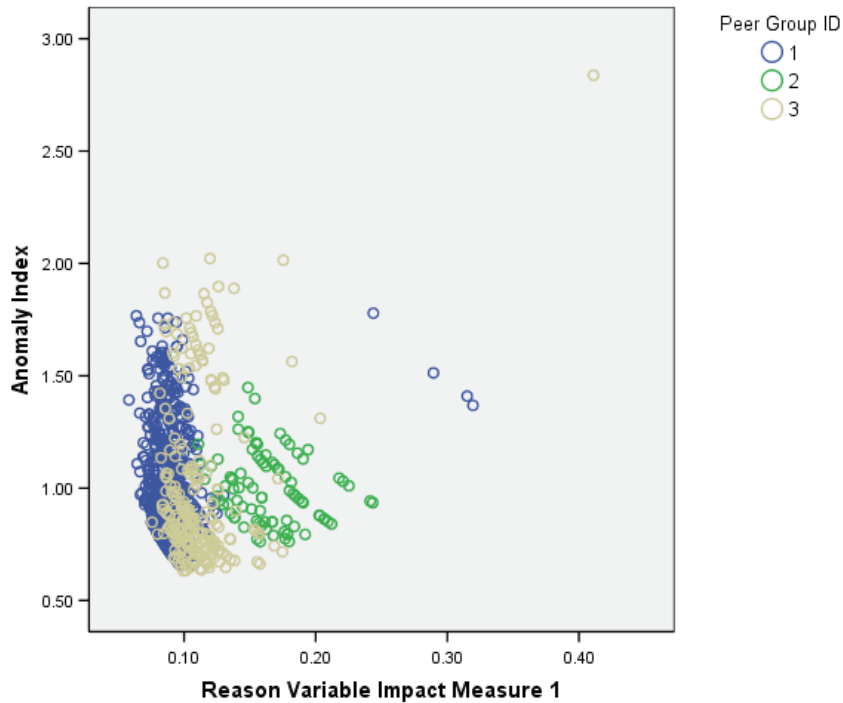
Figure 7-16  
Chart Builder dialog box



- ▶ 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 7-17  
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 cut points.** 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 cut point.

**Selecting cut points.** The potential cut point that produces the greatest information gain is evaluated by the MDLP acceptance criterion. Repeat until no potential cut points are accepted. The accepted cut points 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



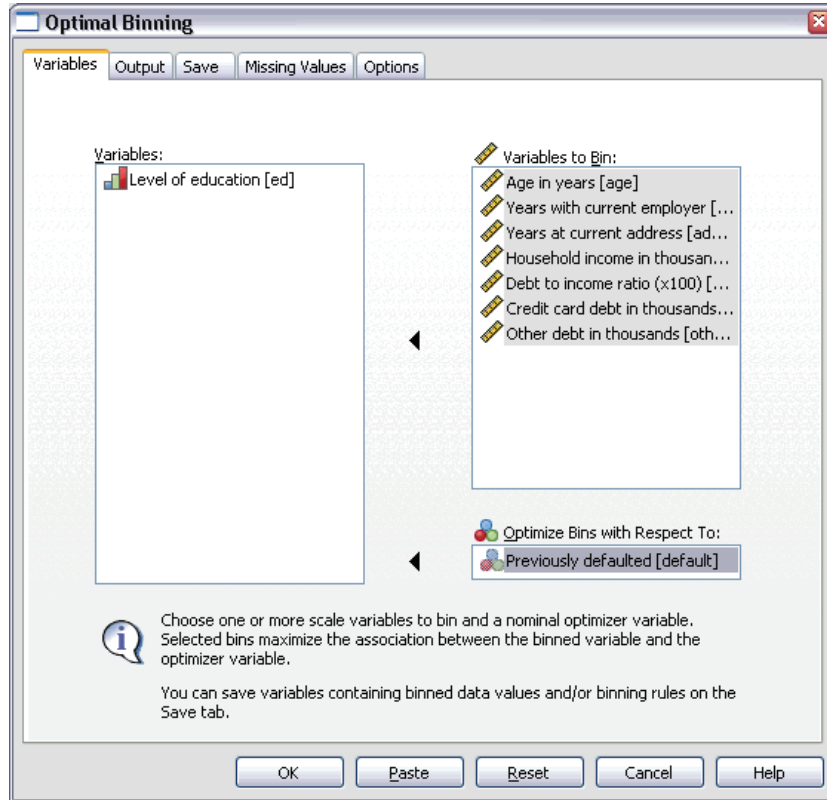
potential predictors are scale, but the loan officer wants to be able to consider models that work best with categorical predictors.

Information on 5000 past customers is collected in *bankloan\_binning.sav*. 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***

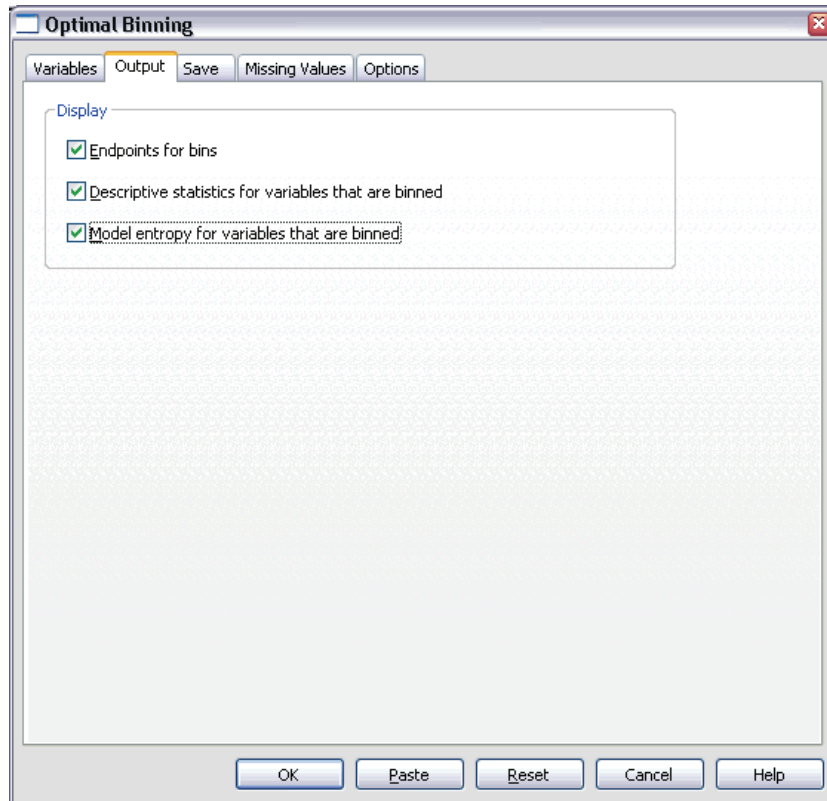
- ▶ To run an Optimal Binning analysis, from the menus choose:  
Transform  
Optimal Binning...

Figure 8-1  
Optimal Binning dialog box, Variables tab



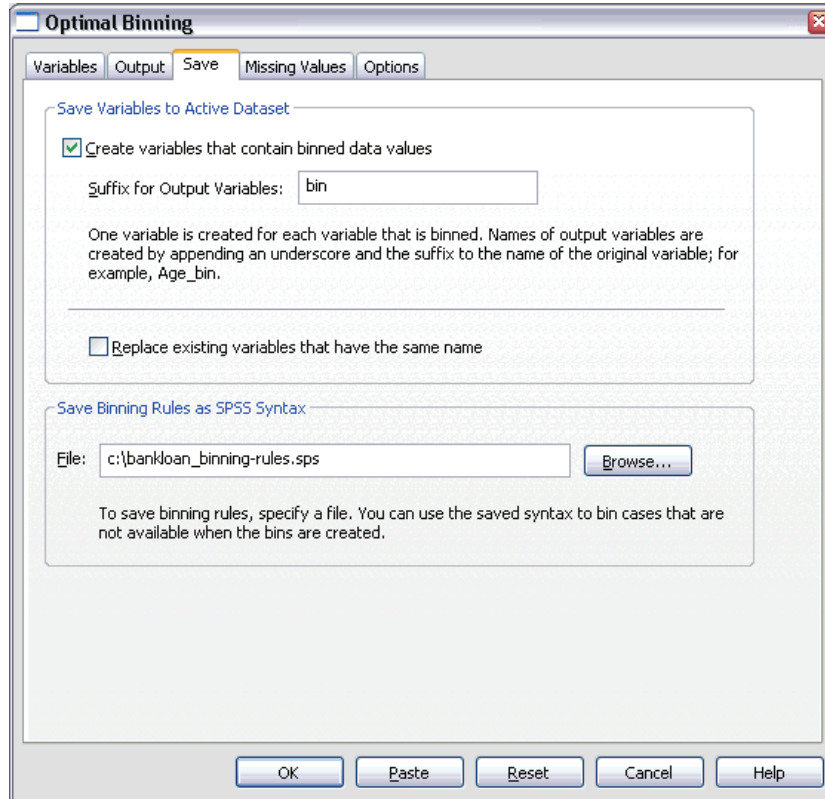
- ▶ Select *Age in years* and *Years with current employer* through *Other debt in thousands* as variables to bin.
- ▶ Select *Previously defaulted* as the guide variable.
- ▶ Click the Output tab.

Figure 8-2  
Optimal Binning dialog box, Output tab



- ▶ Select Descriptive statistics and Model entropy for variables that are binned.
- ▶ Click the Save tab.

Figure 8-3  
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 *c:\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='c:\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.
- SPSS command syntax representing the binning rules is saved to the file *c:\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 8-4  
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 8-5  
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 8-6  
Binning summary for Age in years

Bin	End Point		Number of Cases by Level of Previously defaulted		
	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

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 8-7**  
*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  $\leq$  Household income in thousands  $<$  Upper.

a. Unbounded

The summary for *Household income in thousands* shows a similar pattern, with a single cut point 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 8-8**  
*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  $\leq$  Other debt in thousands  $<$  Upper.

a. Unbounded

The summary for *Other debt in thousands* shows a reversed pattern, with a single cut point 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 8-9**  
*Binning summary for Years with current employer*

Bin	End Point		Number of Cases by Level of Previously defaulted		
	Lower	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

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 8-10**  
*Binning summary for Years at current address*

Bin	End Point		Number of Cases by Level of Previously defaulted		
	Lower	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

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 8-11**  
*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 8-12**  
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  $\leq$  Debt to income ratio ( $\times 100$ )  $<$  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 8-13  
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	

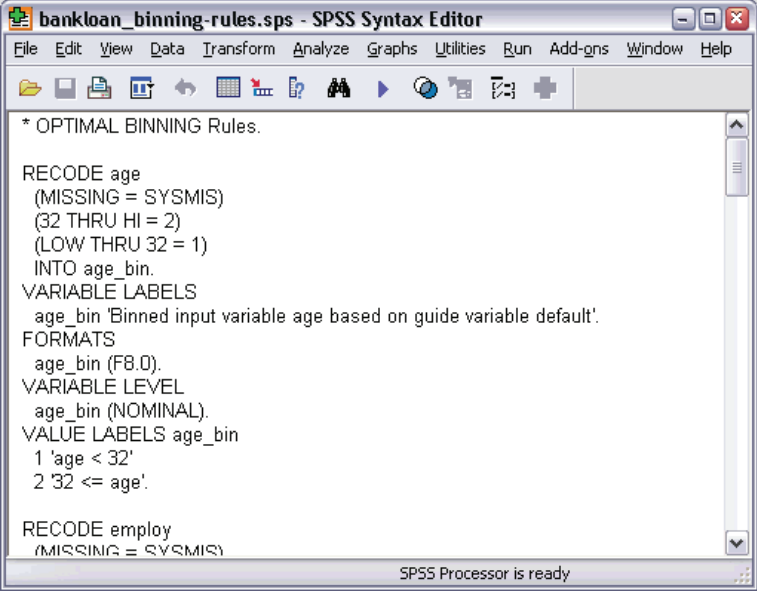
The results of the binning process on this dataset are evident in the Data Editor. These binned variables are useful if you want to produce customized summaries of the binning results using descriptive or reporting procedures, but it is inadvisable to use this dataset to build a predictive model because 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 SPSS command syntax.

- Open *bankloan\_binning-rules.sps*.

Figure 8-14  
Syntax rules file



```
bankloan_binning-rules.sps - SPSS Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Run Add-ons Window Help
* OPTIMAL BINNING Rules.

RECODE age
(MISSING = SYSMIS)
(32 THRU HI = 2)
(LOW THRU 32 = 1)
INTO age_bin.
VARIABLE LABELS
  age_bin 'Binned input variable age based on guide variable default'.
FORMATS
  age_bin (F8.0).
VARIABLE LEVEL
  age_bin (NOMINAL).
VALUE LABELS age_bin
  1 'age < 32'
  2 '32 <= age'.

RECODE employ
(MISSING = SYSMIS)
```

SPSS Processor is ready

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* in the *\Tutorials\sample\_files* subdirectory of the SPSS installation directory, then 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 8-15  
Binned variables for *bankloan.sav* in Data Editor

	preddef3	age bin	employ bin	address bin	income bin	debtinc bin	creddebt bin	othdebt bin
1	.21304	2	3	2	2	2	4	2
2	.43690	1	3	1	2	3	2	2
3	.14102	2	3	3	2	2	1	1
4	.10442	2	3	3	2	1	3	1
5	.43690	1	1	1	2	3	2	2
6	.23358	2	2	1	1	2	1	1
7	.81709	2	4	2	2	4	3	2
8	.11336	2	3	2	2	1	1	1
9	.66390	1	2	1	1	4	2	2
10	.51553	2	1	2	1	4	3	1
11	.09055	1	1	1	1	1	1	1
12	.13631	1	2	1	1	2	1	1
13	.22890	2	4	3	2	2	3	2
14	.40484	2	2	2	2	3	2	2
15	.20866	2	4	3	2	2	3	2

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.

- anomaly indices
  - in Identify Unusual Cases, 23, 25, 71
- binned variables
  - in Optimal Binning, 96
- binning rules
  - in Optimal Binning, 33
- binning summaries
  - in Optimal Binning, 91
- case processing summary
  - in Identify Unusual Cases, 71
- case report
  - in Validate Data, 51, 62
- cross-variable validation rules
  - defining, 54
  - in Define Validation Rules, 8
  - in Validate Data, 15, 61
- data validation
  - in Validate Data, 10
- Define Validation Rules, 4
  - cross-variable rules, 8
  - single-variable rules, 5
- descriptive statistics
  - in Optimal Binning, 89
- duplicate case identifiers
  - in Validate Data, 18, 41
- empty cases
  - in Validate Data, 18
- endpoints for bins
  - in Optimal Binning, 32
- Identify Unusual Cases, 20, 64
  - anomaly case index list, 71
  - anomaly case peer ID list, 72
  - anomaly case reason list, 73
  - anomaly index summary, 78
  - case processing summary, 71
  - categorical variable norms, 76
  - export model file, 25
  - missing values, 26
  - model, 64
  - options, 27
  - output, 23
  - reason summary, 79
  - related procedures, 83
  - save variables, 25
  - scale variable norms, 75
- incomplete case identifiers
  - in Validate Data, 18, 41
- MDLP
  - in Optimal Binning, 30
- missing values
  - in Identify Unusual Cases, 26
- model entropy
  - in Optimal Binning, 90
- Optimal Binning, 30, 84
  - binned variables, 96
  - binning summaries, 91
  - descriptive statistics, 89
  - missing values, 34
  - model, 84
  - model entropy, 90
  - options, 35
  - output, 32
  - save, 33
  - syntax binning rules, 96
- peer group norms
  - in Identify Unusual Cases, 75–76
- peer groups
  - in Identify Unusual Cases, 23, 25, 71–72
- pre-binning
  - in Optimal Binning, 35



## reasons

- in Identify Unusual Cases, 23, 25, 73, 79

## rule descriptions

- in Validate Data, 50

## single-variable validation rules

- defining, 54

- in Define Validation Rules, 5

- in Validate Data, 14

## supervised binning

- in Optimal Binning, 30

- versus unsupervised binning, 30

## unsupervised binning

- versus supervised binning, 30

## Validate Data, 10, 38

- basic checks, 12

- case report, 51, 62

- cross-variable rules, 15, 61

- duplicate case identifiers, 41

- incomplete case identifiers, 41

- output, 16

- related procedures, 63

- rule descriptions, 50

- save variables, 18

- single-variable rules, 14

- variable summary, 50

- warnings, 40

## validation rule violations

- in Validate Data, 18

## validation rules, 3

## variable summary

- in Validate Data, 50

## violations of validation rules

- in Validate Data, 18

## warnings

- in Validate Data, 40