Quick Start Guide

Getting started with IBM SPSS Statistics

Highlights

- Learn two procedures used to investigate and clean datasets: Validate Data and Identify Duplicate Cases.
- View helpful case studies in the SPSS Statistics online Help menu.
- Find public training courses for additional skill development.

Introduction

IBM® SPSS® Statistics is a family of integrated software products for general-purpose data analysis including reporting, graphing and statistics.

Many spreadsheet users migrate to SPSS Statistics because the sheer quantity of data is too large for a spreadsheet. Others have concerns regarding the time required to create and maintain complex spreadsheets, or they may have quality assurance issues. IBM SPSS Statistics has a wide variety of built-in reporting and statistical analysis procedures developed by statisticians and is professionally tested to ensure the quality of the resulting calculations.

This guide introduces two procedures used to investigate and clean datasets: *Validate Data* and *Identify Duplicate Cases*.

Validate Data

Step 1

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

- Open data file telco_missing.sav. This file is located in ProgramFiles/IBM/SPSS/23/Samples/English on your C:/ or program installation drive.
- In the Statistics Data Editor window menu click *Data*.
- Go to *Validation*.
- Click *Load Predefined Rules*. [Note: This is appropriate if you are a new user of this procedure. Statistics will display a message warning you that any previously defined rules will be replaced. Ignore this message, since you have not defined any custom rules. IBM SPSS Statistics will pre-load some standard, common practice rules.]
- Click *OK*.
- Again, in the Statistics Data Editor menu, click *Data*.
- Select *Validation*.
- Click *Validate Data* (Figure 1).

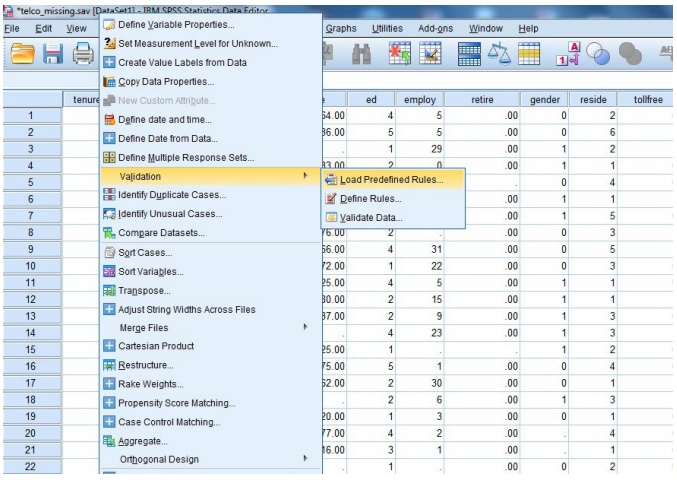

Figure 1: Selecting Validate Data in the Statistics menu.

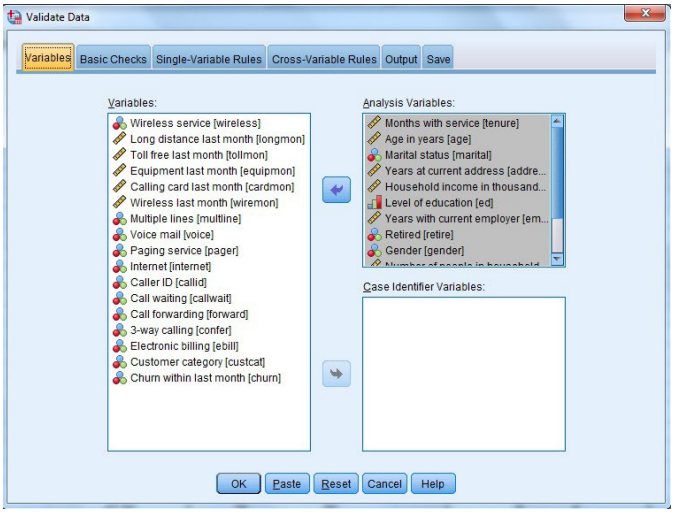

Figure 2: Setting up the Validate Data procedure: The variables *Months with service* through *Number of people in household* from telco_missing.sav have been selected and placed in the Analysis Variables box within the Validate Data dialog window.

Step 2

A selection of predefined rules is provided for your use, and you can add additional rules as needed—simple or complex.

- Select all variables that you wish to validate from the list on the left.
- Move these to the Analysis Variables list (Figure 2).
- Click the *Single-Variable Rules* tab.
- This tab displays summary statistics for the analysis variables and enables you to specify validation rules based on individual variables.
- Click the *Define Rules* button to modify or add new rules. Look over this tab, but don't set up a rule.
- Click *Continue*.

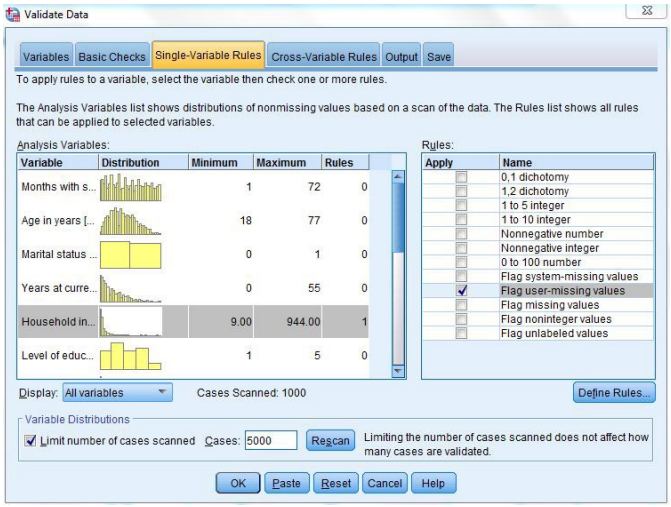

Figure 3: Applying a rule to a selected variable: Under the Single-Variable Rules tab of the Validate Data dialog box, the variable, *Household income in thousands* has been selected in the list of Analysis Variables. The *Flag user-missing values* has been selected in the Rules list.

Step 2 (continued)

Select a variable from the Analysis Variable list.

- From the rules list, check the box(es) to apply rule(s) to the selected variable(s) (Figure 3).
- Click the *Cross-Variable Rules* tab. [Note: This tab is used to define rules that involve more than one variable. The logical expression defines variable values that may appear valid individually, but are invalid in combination. A classic example of this is the case which indicates gender: male and pregnancy test result: positive. For now, look over this tab but don't set up a rule.]
- Click the *Define Rules* button.
- Click *Continue*.
- Click *OK*.

Step 3

The Validate Data procedure creates several tables which summarize rule violations in different ways. If none of the data violates any of the rules in use, a message will be generated explaining that no rules were violated.

- The Variable Summary is organized by the analysis variables.
- For each variable, there is a list of the rules violated.
- The final column gives the number of times the rule was violated for that specific variable (Figure 4).

Variable Summary

	Rule	Number of Violations
Household income in thousands	Flag missing values	179
	Total	179

Case Report[b]

Case	Validation Rule Violations: Single-Variable[a]
3	Flag missing values (1)
14	Flag missing values (1)
18	Flag missing values (1)
22	Flag missing values (1)
26	Flag missing values (1)
32	Flag missing values (1)
40	Flag missing values (1)
51	Flag missing values (1)
58	Flag missing values (1)
59	Flag missing values (1)

Figure 4: Variable Summary and Case Report: The Output window provides a summary and the individual cases which have been identified by the Rules in the Validate Data dialog window.

Identify Duplicate Cases

Step 1

The Identify Duplicate Cases procedure enables you to identify cases that have duplicate values based on any combination of variables selected.

- Open the file stroke_invalid.sav. This file is located in the directory of a standard SPSS Statistics installation.
- In the Statistics Data Editor window menu click *Data*.
- Click *Identify Duplicate Cases* (Figure 5).

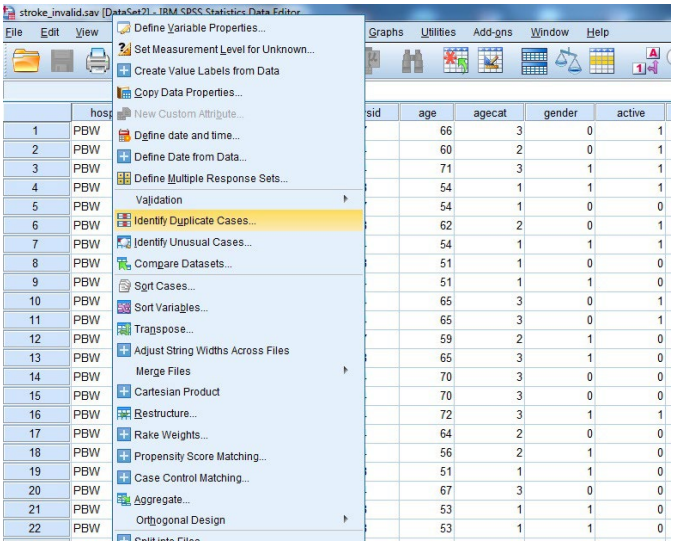

Figure 5: Selecting Identify Duplicate Cases in the Statistics menu.

Step 2

You can specify rules for automatic identification of primary cases. For example, you might want to identify the case with the most recent creation date as primary. Earlier cases in the same group would be duplicates.

- Select variables that define a duplicate case from the list on the left.
- To search for exact duplicates—cases that are identical in every field—select all variables. In most cases, duplicates are defined using only a subset of variables in the dataset.
- Move these to the *Define Matching Cases By* list (Figure 6).
- Click *OK*.

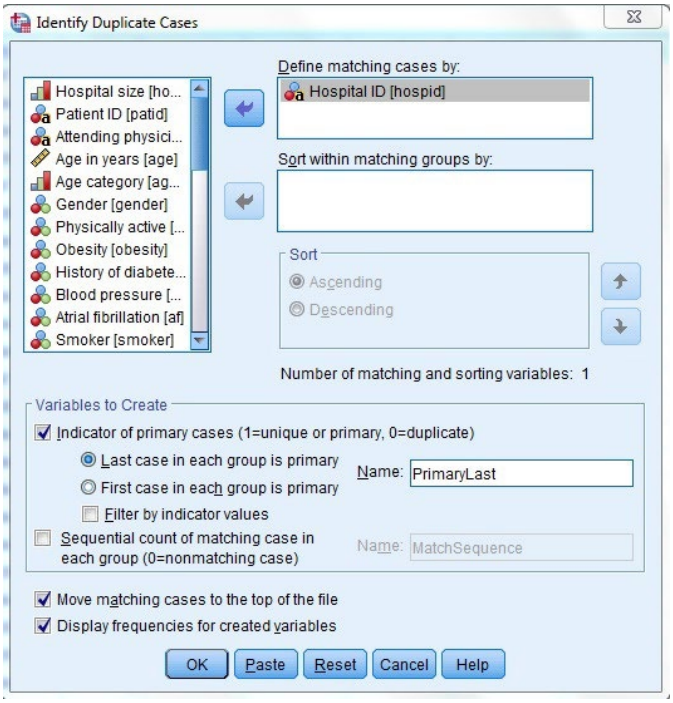

Figure 6: Setting up the Identify Duplicate Cases procedure: A number of variables from stroke_invalid.sav has been selected and placed in the Define matching cases by box within the Identify Duplicate Cases dialog window.

Step 3

You can see how much of your dataset is duplicate cases.

- In the Output window, the *Indicator of each last matching case as Primary* table summarizes the results (Figure 7).
- This is a frequency table which indicates how many cases were primary, and how many were duplicates.
- The default definition of a primary case is the last case in a group of cases whose values match. You can change this definition to use the first case instead, and you can use other variables to sort the groups for more flexibility in defining the primary case.

Statistics

N	Valid	1183
	Missing	0

Indicator of each last matching case as Primary

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Duplicate Case	1167	98.6	98.6	98.6
	Primary Case	16	1.4	1.4	100.0
	Total	1183	100.0	100.0	

Figure 7: Indicator of each last matching case as primary: The Output window provides a summary of primary and duplicate cases which have been identified by Define matching cases within the Identify Duplicate Cases dialog window.

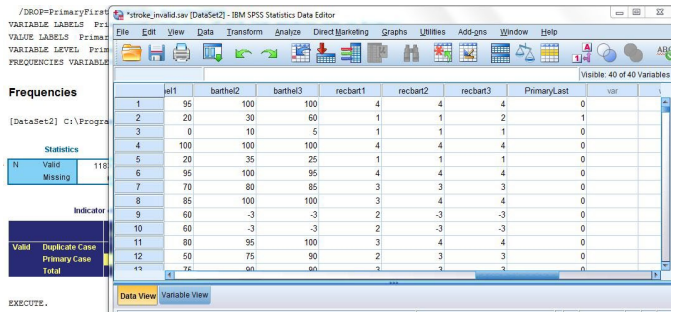

Figure 8: Duplicate cases indicated in the Data Editor.

Step 4

An indicator variable is useful for further data manipulation.

- A new variable is created in the dataset, which indicates whether the case was primary or a duplicate (Figure 8). [This variable can be saved with the dataset. Most often, this indicator variable will be used for filtering the data for analysis, or to delete duplicate cases.]

IBM SPSS Statistics case studies

Validate Data

In the IBM SPSS Statistics Menu, select *Help…Case Studies*. An outline of the Case Studies will open. Expand *Data Preparation Option* and select *Validate Data*.

Identify Duplicate Cases

In the IBM SPSS Statistics Menu, select *Help…Case Studies*. An outline of the Case Studies will open. Expand *Data Preparation Option*, expand *Validate Data* and click on *Validating a Medical Database*.

IBM SPSS public training courses

Introduction to IBM SPSS Statistics

A two-day course available as live web-based training. The course guides you through the fundamentals of using IBM SPSS Statistics for the typical data analysis process. Learn the basics of reading data, data definition, data modification, data analysis and presentation of your results. See how easy it is to get your data into IBM SPSS Statistics so that you can focus on analyzing the information. In addition to the fundamentals, learn shortcuts that will help you save time.

Data management and manipulation with IBM SPSS Statistics

A two-day course available as live web-based training. Use a wide range of transformation techniques to modify data values and discover how to automate your work, manipulate your data files and results, and export your results to other applications' file formats. You will also gain an understanding of the various options for controlling the SPSS Statistics operating environment and learn how to use basic syntax to perform data transformations efficiently.

For more information

Check the IBM SPSS training website or contact your IBM SPSS representative at 800.543.2185 for more details.

IBM Corporation
IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
May 2015

YTY03002-USEN-03