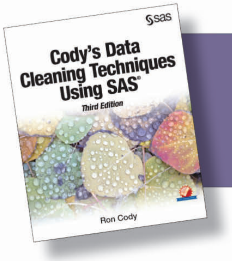

From *Cody's Data Cleaning Techniques Using SAS®, Third Edition.*
Full book available for purchase here.

Contents

List of Programs ix

About This Book xv

About The Author xvii

Acknowledgments xix

Introduction xxi

Chapter 1: Working with Character Data 1

Introduction 1

Using PROC FREQ to Detect Character Variable Errors 4

Changing the Case of All Character Variables in a Data Set 6

A Summary of Some Character Functions (Useful for Data Cleaning) 8

- UPCASE, LOWCASE, and PROPCASE 8
- NOTDIGIT, NOTALPHA, and NOTALNUM 8
- VERIFY 9
- COMPBL 9
- COMPRESS 9
- MISSING 10
- TRIMN and STRIP 10

Checking that a Character Value Conforms to a Pattern 11

Using a DATA Step to Detect Character Data Errors 12

Using PROC PRINT with a WHERE Statement to Identify Data Errors 13

Using Formats to Check for Invalid Values 14

Creating Permanent Formats 16

Removing Units from a Value 17

Removing Non-Printing Characters from a Character Value 18

Conclusions 19

Chapter 2: Using Perl Regular Expressions to Detect Data Errors 21

Introduction 21

Describing the Syntax of Regular Expressions 21

Checking for Valid ZIP Codes and Canadian Postal Codes 23

Searching for Invalid Email Addresses 25

Verifying Phone Numbers 26

Converting All Phone Numbers to a Standard Form 27
Developing a Macro to Test Regular Expressions 28
Conclusions 29
Chapter 3: Standardizing Data 31
Introduction 31
Using Formats to Standardize Company Names 31
Creating a Format from a SAS Data Set 33
Using TRANWRD and Other Functions to Standardize Addresses 36
Using Regular Expressions to Help Standardize Addresses 38
Performing a "Fuzzy" Match between Two Files 40
Conclusions 44
Chapter 4: Data Cleaning Techniques for Numeric Data 45
Introduction 45
Using PROC UNIVARIATE to Examine Numeric Variables 45
Describing an ODS Option to List Selected Portions of the Output 49
Listing Output Objects Using the Statement TRACE ON 52
Using a PROC UNIVARIATE Option to List More Extreme Values 52
Presenting a Program to List the 10 Highest and Lowest Values 53
Presenting a Macro to List the *n* Highest and Lowest Values 55
Describing Two Programs to List the Highest and Lowest Values by Percentage 58
    Using PROC UNIVARIATE 58
    Presenting a Macro to List the Highest and Lowest *n*% Values 60
    Using PROC RANK 62
Using Pre-Determined Ranges to Check for Possible Data Errors 66
Identifying Invalid Values versus Missing Values 67
Checking Ranges for Several Variables and Generating a Single Report 69
Conclusions 72
Chapter 5: Automatic Outlier Detection for Numeric Data 73
Introduction 73
Automatic Outlier Detection (Using Means and Standard Deviations) 73
Detecting Outliers Based on a Trimmed Mean and Standard Deviation 75
Describing a Program that Uses Trimmed Statistics for Multiple Variables 78
Presenting a Macro Based on Trimmed Statistics 81
Detecting Outliers Based on the Interquartile Range 83
Conclusions 86

Chapter 6: More Advanced Techniques for Finding Errors in Numeric Data 87

Introduction 87

Introducing the Banking Data Set 87

Running the Auto_Outliers Macro on Bank Deposits 91

Identifying Outliers Within Each Account 92

Using Box Plots to Inspect Suspicious Deposits 95

Using Regression Techniques to Identify Possible Errors in the Banking Data 99

Using Regression Diagnostics to Identify Outliers 104

Conclusions 108

Chapter 7: Describing Issues Related to Missing and Special Values (Such as 999) 109

Introduction 109

Inspecting the SAS Log 109

Using PROC MEANS and PROC FREQ to Count Missing Values 110

Counting Missing Values for Numeric Variables 110

Counting Missing Values for Character Variables 111

Using DATA Step Approaches to Identify and Count Missing Values 113

Locating Patient Numbers for Records Where Patno Is Either Missing or Invalid 113

Searching for a Specific Numeric Value 117

Creating a Macro to Search for Specific Numeric Values 119

Converting Values Such as 999 to a SAS Missing Value 121

Conclusions 121

Chapter 8: Working with SAS Dates 123

Introduction 123

Changing the Storage Length for SAS Dates 123

Checking Ranges for Dates (Using a DATA Step) 124

Checking Ranges for Dates (Using PROC PRINT) 125

Checking for Invalid Dates 125

Working with Dates in Nonstandard Form 128

Creating a SAS Date When the Day of the Month Is Missing 129

Suspending Error Checking for Known Invalid Dates 131

Conclusions 131

Chapter 9: Looking for Duplicates and Checking Data with Multiple Observations per Subject 133

Introduction 133

Eliminating Duplicates by Using PROC SORT 133

Demonstrating a Possible Problem with the NODUPRECS Option 136
Reviewing First. and Last. Variables 138
Detecting Duplicates by Using DATA Step Approaches 140
Using PROC FREQ to Detect Duplicate IDs 141
Working with Data Sets with More Than One Observation per Subject 143
Identifying Subjects with *n* Observations Each (DATA Step Approach) 144
Identifying Subjects with *n* Observations Each (Using PROC FREQ) 146
Conclusions 146

Chapter 10: Working with Multiple Files 147

Introduction 147
Checking for an ID in Each of Two Files 147
Checking for an ID in Each of *n* Files 150
A Macro for ID Checking 152
Conclusions 154

Chapter 11: Using PROC COMPARE to Perform Data Verification 155

Introduction 155
Conducting a Simple Comparison of Two Data Files 155
Simulating Double Entry Verification Using PROC COMPARE 160
Other Features of PROC COMPARE 161
Conclusions 162

Chapter 12: Correcting Errors 163

Introduction 163
Hard Coding Corrections 163
Describing Named Input 164
Reviewing the UPDATE Statement 166
Using the UPDATE Statement to Correct Errors in the Patients Data Set 168
Conclusions 171

Chapter 13: Creating Integrity Constraints and Audit Trails 173

Introduction 173
Demonstrating General Integrity Constraints 174
Describing PROC APPEND 177
Demonstrating How Integrity Constraints Block the Addition of Data Errors 178
Adding Your Own Messages to Violations of an Integrity Constraint 179
Deleting an Integrity Constraint Using PROC DATASETS 180
Creating an Audit Trail Data Set 180

Demonstrating an Integrity Constraint Involving More Than One Variable 183
Demonstrating a Referential Constraint 186
Attempting to Delete a Primary Key When a Foreign Key Still Exists 188
Attempting to Add a Name to the Child Data Set 190
Demonstrating How to Delete a Referential Constraint 191
Demonstrating the CASCADE Feature of a Referential Constraint 191
Demonstrating the SET NULL Feature of a Referential Constraint 192
Conclusions 193
Chapter 14: A Summary of Useful Data Cleaning Macros 195
Introduction 195
A Macro to Test Regular Expressions 195
A Macro to List the *n* Highest and Lowest Values of a Variable 196
A Macro to List the *n*% Highest and Lowest Values of a Variable 197
A Macro to Perform Range Checks on Several Variables 198
A Macro that Uses Trimmed Statistics to Automatically Search for Outliers 200
A Macro to Search a Data Set for Specific Values Such as 999 202
A Macro to Check for ID Values in Multiple Data Sets 203
Conclusions 204
Index 205

From Cody's Data Cleaning Techniques Using SAS®, Third Edition by Ron Cody. Copyright © 2017, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.