## Chapter 3: Standardizing Data

### Introduction

Using Formats to Standardize Company Names
Creating a Format from a SAS Data Set
Using TRANWRD and Other Functions to Standardize Addresses
Using Regular Expressions to Help Standardize Addresses
Performing a "Fuzzy" Match between Two Files

### Conclusions

---

## Chapter 4: Data Cleaning Techniques for Numeric Data

### Introduction

Using PROC UNIVARIATE to Examine Numeric Variables
Describing an ODS Option to List Selected Portions of the Output
Listing Output Objects Using the Statement TRACE ON
Using a PROC UNIVARIATE Option to List More Extreme Values
Presenting a Program to List the 10 Highest and Lowest Values
Describing Two Programs to List the Highest and Lowest Values by Percentage
Using PROC UNIVARIATE
Presenting a Macro to List the Highest and Lowest n% Values
Using PROC RANK
Using Pre-Determined Ranges to Check for Possible Data Errors
Identifying Invalid Values versus Missing Values
Checking Ranges for Several Variables and Generating a Single Report

### Conclusions

---

## Chapter 5: Automatic Outlier Detection for Numeric Data

### Introduction

Automatic Outlier Detection (Using Means and Standard Deviations)
Detecting Outliers Based on a Trimmed Mean and Standard Deviation
Describing a Program that Uses Trimmed Statistics for Multiple Variables
Presenting a Macro Based on Trimmed Statistics
Detecting Outliers Based on the Interquartile Range

### Conclusions

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

- **Introduction**
- **Introducing the Banking Data Set**
- **Running the Auto_Outliers Macro on Bank Deposits**
- **Identifying Outliers Within Each Account**
- **Using Box Plots to Inspect Suspicious Deposits**
- **Using Regression Techniques to Identify Possible Errors in the Banking Data**
- **Using Regression Diagnostics to Identify Outliers**
- **Conclusions**

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

- **Introduction**
- **Inspecting the SAS Log**
- **Using PROC MEANS and PROC FREQ to Count Missing Values**
  - Counting Missing Values for Numeric Variables
  - Counting Missing Values for Character Variables
- **Using DATA Step Approaches to Identify and Count Missing Values**
- **Locating Patient Numbers for Records Where Patno Is Either Missing or Invalid**
- **Searching for a Specific Numeric Value**
- **Creating a Macro to Search for Specific Numeric Values**
- **Converting Values Such as 999 to a SAS Missing Value**
- **Conclusions**

**Chapter 8: Working with SAS Dates**

- **Introduction**
- **Changing the Storage Length for SAS Dates**
- **Checking Ranges for Dates (Using a DATA Step)**
- **Checking Ranges for Dates (Using PROC PRINT)**
- **Checking for Invalid Dates**
- **Working with Dates in Nonstandard Form**
- **Creating a SAS Date When the Day of the Month Is Missing**
- **Suspending Error Checking for Known Invalid Dates**
- **Conclusions**

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

- **Introduction**
- **Eliminating Duplicates by Using PROC SORT**
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