

A pair of metal dividers and a scriber are shown on a textured, aged surface. The dividers are positioned diagonally, with their legs spread. The scriber lies horizontally across the middle of the dividers. The lighting is warm, creating a golden-brown hue and casting soft shadows.

Mamdouh Refaat

Data  
Preparation  
*for* Data Mining  
Using SAS

# **DATA PREPARATION FOR DATA MINING USING SAS**

## The Morgan Kaufmann Series in Data Management Systems

*Series Editor:* Jim Gray, Microsoft Research

*Data Preparation for Data Mining Using SAS*  
Mamdouh Refaat

*Querying XML: XQuery, XPath, and SQL/XML in Context*

Jim Melton and Stephen Buxton

*Data Mining: Concepts and Techniques*,  
Second Edition

Jiawei Han and Micheline Kamber

*Database Modeling and Design: Logical Design*,  
Fourth Edition

Toby J. Teorey, Sam S. Lightstone and Thomas P. Nadeau

*Foundations of Multidimensional and Metric Data Structures*

Hanan Samet

*Joe Celko's SQL for Smarties: Advanced SQL Programming*, Third Edition

Joe Celko

*Moving Objects Databases*

Ralf Hartmut Güting and Markus Schneider

*Joe Celko's SQL Programming Style*

Joe Celko

*Data Mining, Second Edition: Concepts and Techniques*

Ian Witten and Eibe Frank

*Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*

Earl Cox

*Data Modeling Essentials*, Third Edition

Graeme C. Simsion and Graham C. Witt

*Location-Based Services*

Jochen Schiller and Agnès Voisard

*Database Modeling with Microsoft® Visio for Enterprise Architects*

Terry Halpin, Ken Evans, Patrick Hallock, Bill Maclean

*Designing Data-Intensive Web Applications*

Stephano Ceri, Piero Fraternali, Aldo Bongio, Marco Brambilla, Sara Comai, and Maristella Matera

*Mining the Web: Discovering Knowledge from Hypertext Data*

Soumen Chakrabarti

*Advanced SQL: 1999—Understanding Object-Relational and Other Advanced Features*

Jim Melton

*Database Tuning: Principles, Experiments, and Troubleshooting Techniques*

Dennis Shasha and Philippe Bonnet

*SQL:1999—Understanding Relational Language Components*

Jim Melton and Alan R. Simon

*Information Visualization in Data Mining and Knowledge Discovery*

Edited by Usama Fayyad, Georges G. Grinstein, and Andreas Wierse

*Transactional Information Systems: Theory, Algorithms, and Practice of Concurrency Control and Recovery*

Gerhard Weikum and Gottfried Vossen

*Spatial Databases: With Application to GIS*

Philippe Rigaux, Michel Scholl, and Agnes Voisard

*Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design*

Terry Halpin

*Component Database Systems*

Edited by Klaus R. Dittrich and Andreas Geppert

*Managing Reference Data in Enterprise Databases: Binding Corporate Data to the Wider World*

Malcolm Chisholm

*Understanding SQL and Java Together: A Guide to SQLJ, JDBC, and Related Technologies*

Jim Melton and Andrew Eisenberg

*Database: Principles, Programming, and Performance*, Second Edition

Patrick and Elizabeth O'Neil

*The Object Data Standard: ODMG 3.0*

Edited by R. G. G. Cattell and Douglas K. Barry

*Data on the Web: From Relations to Semistructured Data and XML*

Serge Abiteboul, Peter Buneman, and Dan Suciu

*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*

Ian Witten and Eibe Frank

*Joe Celko's SQL for Smarties: Advanced SQL Programming*, Second Edition

Joe Celko

*Joe Celko's Data and Databases: Concepts in Practice*

Joe Celko

*Developing Time-Oriented Database Applications in SQL*

Richard T. Snodgrass

*Web Farming for the Data Warehouse*

Richard D. Hackathorn

*Management of Heterogeneous and Autonomous Database Systems*

Edited by Ahmed Elmagarmid, Marek Rusinkiewicz, and Amit Sheth

*Object-Relational DBMSs: Tracking the Next Great Wave*, Second Edition

Michael Stonebraker and Paul Brown, with Dorothy Moore

*A Complete Guide to DB2 Universal Database*

Don Chamberlin

*Universal Database Management: A Guide to Object/Relational Technology*

Cynthia Maro Saracco

*Readings in Database Systems*, Third Edition

Edited by Michael Stonebraker and

Joseph M. Hellerstein

*Understanding SQL's Stored Procedures: A Complete Guide to SQL/PSM*

Jim Melton

*Principles of Multimedia Database Systems*

V. S. Subrahmanian

*Principles of Database Query Processing for Advanced Applications*

Clement T. Yu and Weiyi Meng

*Advanced Database Systems*

Carlo Zaniolo, Stefano Ceri, Christos

Faloutsos, Richard T. Snodgrass, V. S.

Subrahmanian, and Roberto Zicari

*Principles of Transaction Processing*

Philip A. Bernstein and Eric Newcomer

*Using the New DB2: IBM's Object-Relational Database System*

Don Chamberlin

*Distributed Algorithms*

Nancy A. Lynch

*Active Database Systems: Triggers and Rules For Advanced Database Processing*

Edited by Jennifer Widom and Stefano Ceri

*Migrating Legacy Systems: Gateways, Interfaces, & the Incremental Approach*

Michael L. Brodie and Michael Stonebraker

*Atomic Transactions*

Nancy Lynch, Michael Merritt, William Weihl, and Alan Fekete

*Query Processing for Advanced Database Systems*

Edited by Johann Christoph Freytag, David Maier, and Gottfried Vossen

*Transaction Processing: Concepts and Techniques*

Jim Gray and Andreas Reuter

*Building an Object-Oriented Database System: The Story of O<sub>2</sub>*

Edited by François Bancilhon, Claude Delobel, and Paris Kanellakis

*Database Transaction Models for Advanced Applications*

Edited by Ahmed K. Elmagarmid

*A Guide to Developing Client/Server SQL Applications*

Setrag Khoshafian, Arvola Chan, Anna Wong, and Harry K. T. Wong

*The Benchmark Handbook for Database and Transaction Processing Systems*, Second Edition

Edited by Jim Gray

*Camelot and Avalon: A Distributed Transaction Facility*

Edited by Jeffrey L. Eppinger, Lily B.

Mummert, and Alfred Z. Spector

*Readings in Object-Oriented Database Systems*

Edited by Stanley B. Zdonik and David Maier

# DATA PREPARATION FOR DATA MINING USING SAS

MAMDOUH REFAAT



AMSTERDAM • BOSTON • HEIDELBERG • LONDON  
NEW YORK • OXFORD • PARIS • SAN DIEGO  
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO  
Morgan Kaufmann Publishers is an imprint of Elsevier



*Publisher:* Diane D. Cerra  
*Assistant Editor:* Asma Palmeiro  
*Publishing Services Manager:* George Morrison  
*Project Manager:* Marilyn E. Rash  
*Cover Design:* Eric DeCicco  
*Cover Image:* Qettyimages  
*Composition:* diacriTech  
*Copyeditor:* Joan Flaherty  
*Proofreader:* Dianne Wood  
*Indexer:* Ted Laux  
*Interior printer:* The Maple Press Company  
*Cover printer:* Phoenix Color Corp.

Morgan Kaufmann Publishers is an imprint of Elsevier.  
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

© 2007 by Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan Kaufmann Publishers is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

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, scanning, or otherwise—without prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: [permissions@elsevier.com](mailto:permissions@elsevier.com). You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

#### **Library of Congress Cataloging-in-Publication Data**

Refaat, Mamdouh.

Data preparation for data mining using SAS / Mamdouh Refaat.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-12-373577-5 (alk. paper)

ISBN-10: 0-12-373577-7 (alk. paper)

1. Data mining. 2. SAS (Computer file) 1. Title.

QA76.9.D343R43 2007

006.3'12—dc22

2006023681

ISBN 13: 978-0-12-373577-5

ISBN 10: 0-12-373577-7

For information on all Morgan Kaufmann publications, visit our  
Web site at [www.mkp.com](http://www.mkp.com) or [www.books.elsevier.com](http://www.books.elsevier.com)

Printed in the United States of America

06 07 08 09 10 5 4 3 2 1

Working together to grow  
libraries in developing countries

[www.elsevier.com](http://www.elsevier.com) | [www.bookaid.org](http://www.bookaid.org) | [www.sabre.org](http://www.sabre.org)

ELSEVIER

BOOK AID  
International

Sabre Foundation

# CONTENTS

<b>LIST OF FIGURES</b>	xv
<b>LIST OF TABLES</b>	xvii
<b>PREFACE</b>	xxi

## CHAPTER

# 1

<b>INTRODUCTION</b>	1
---------------------	---

---

1.1 The Data Mining Process	1
1.2 Methodologies of Data Mining	1
1.3 The Mining View	3
1.4 The Scoring View	4
1.5 Notes on Data Mining Software	4

## CHAPTER

# 2

<b>TASKS AND DATA FLOW</b>	7
----------------------------	---

---

2.1 Data Mining Tasks	7
2.2 Data Mining Competencies	9
2.3 The Data Flow	10
2.4 Types of Variables	11
2.5 The Mining View and the Scoring View	12
2.6 Steps of Data Preparation	13

## CHAPTER

# 3

<b>REVIEW OF DATA MINING MODELING TECHNIQUES</b>	15
--	----

---

3.1 Introduction	15
3.2 Regression Models	15
3.2.1 Linear Regression	16

3.2.2 Logistic Regression	18
3.3 Decision Trees	21
3.4 Neural Networks	22
3.5 Cluster Analysis	25
3.6 Association Rules	26
3.7 Time Series Analysis	26
3.8 Support Vector Machines	26

## CHAPTER

## 4

<b>SAS MACROS: A QUICK START</b>	<b>29</b>
----------------------------------	-----------

---

4.1 Introduction: Why Macros?	29
4.2 The Basics: The Macro and Its Variables	30
4.3 Doing Calculations	32
4.4 Programming Logic	33
4.5 Working with Strings	35
4.6 Macros That Call Other Macros	36
4.7 Common Macro Patterns and Caveats	37
4.7.1 Generating a List of Macro Variables	37
4.7.2 Double Coding	39
4.7.3 Using Local Variables	39
4.7.4 From a DATA Step to Macro Variables	40
4.8 Where to Go From Here	41

## CHAPTER

## 5

<b>DATA ACQUISITION AND INTEGRATION</b>	<b>43</b>
---	-----------

---

5.1 Introduction	43
5.2 Sources of Data	43
5.2.1 Operational Systems	43
5.2.2 Data Warehouses and Data Marts	44
5.2.3 OLAP Applications	44
5.2.4 Surveys	44
5.2.5 Household and Demographic Databases	45
5.3 Variable Types	45
5.3.1 Nominal Variables	45
5.3.2 Ordinal Variables	46
5.3.3 Real Measures	47

5.4	Data Rollup	47
5.5	Rollup with Sums, Averages, and Counts	54
5.6	Calculation of the Mode	55
5.7	Data Integration	56
5.7.1	Merging	57
5.7.2	Concatenation	59
<b>CHAPTER 6</b>	<b>INTEGRITY CHECKS</b>	<b>63</b>
6.1	Introduction	63
6.2	Comparing Datasets	66
6.3	Dataset Schema Checks	66
6.3.1	Dataset Variables	66
6.3.2	Variable Types	69
6.4	Nominal Variables	70
6.4.1	Testing the Presence of All Categories	70
6.4.2	Testing the Similarity of Ratios	73
6.5	Continuous Variables	76
6.5.1	Comparing Measure from Two Datasets	77
6.5.2	Comparing the Means, Standard Deviations, and Variance	78
6.5.3	The Confidence-Level Calculations Assumptions	80
6.5.4	Comparison of Other Measures	81
<b>CHAPTER 7</b>	<b>EXPLORATORY DATA ANALYSIS</b>	<b>83</b>
7.1	Introduction	83
7.2	Common EDA Procedures	83
7.3	Univariate Statistics	84
7.4	Variable Distribution	86
7.5	Detection of Outliers	86
7.5.1	Identification of Outliers Using Ranges	88
7.5.2	Identification of Outliers Using Model Fitting	91
7.5.3	Identification of Outliers Using Clustering	94
7.5.4	Notes on Outliers	96
7.6	Testing Normality	96
7.7	Cross-tabulation	97
7.8	Investigating Data Structures	97



## CHAPTER

**8****SAMPLING AND PARTITIONING**

99

---

8.1	Introduction	99
8.2	Contents of Samples	100
8.3	Random Sampling	101
8.3.1	Constraints on Sample Size	101
8.3.2	SAS Implementation	101
8.4	Balanced Sampling	104
8.4.1	Constraints on Sample Size	105
8.4.2	SAS Implementation	106
8.5	Minimum Sample Size	110
8.5.1	Continuous and Binary Variables	110
8.5.2	Sample Size for a Nominal Variable	112
8.6	Checking Validity of Sample	113

## CHAPTER

**9****DATA TRANSFORMATIONS**

115

---

9.1	Raw and Analytical Variables	115
9.2	Scope of Data Transformations	116
9.3	Creation of New Variables	119
9.3.1	Renaming Variables	120
9.3.2	Automatic Generation of Simple Analytical Variables	124
9.4	Mapping of Nominal Variables	126
9.5	Normalization of Continuous Variables	130
9.6	Changing the Variable Distribution	131
9.6.1	Rank Transformations	131
9.6.2	Box–Cox Transformations	133
9.6.3	Spreading the Histogram	138

## CHAPTER

**10****BINNING AND REDUCTION  
OF CARDINALITY**

141

---

10.1	Introduction	141
10.2	Cardinality Reduction	142
10.2.1	The Main Questions	142

10.2.2	Structured Grouping Methods	144
10.2.3	Splitting a Dataset	144
10.2.4	The Main Algorithm	145
10.2.5	Reduction of Cardinality Using Gini Measure	147
10.2.6	Limitations and Modifications	156
10.3	Binning of Continuous Variables	157
10.3.1	Equal-Width Binning	157
10.3.2	Equal-Height Binning	160
10.3.3	Optimal Binning	164

## CHAPTER

## 11

## TREATMENT OF MISSING VALUES 171

---

11.1	Introduction	171
11.2	Simple Replacement	174
11.2.1	Nominal Variables	174
11.2.2	Continuous and Ordinal Variables	176
11.3	Imputing Missing Values	179
11.3.1	Basic Issues in Multiple Imputation	179
11.3.2	Patterns of Missingness	180
11.4	Imputation Methods and Strategy	181
11.5	SAS Macros for Multiple Imputation	185
11.5.1	Extracting the Pattern of Missing Values	185
11.5.2	Reordering Variables	190
11.5.3	Checking Missing Pattern Status	194
11.5.4	Imputing to a Monotone Missing Pattern	197
11.5.5	Imputing Continuous Variables	198
11.5.6	Combining Imputed Values of Continuous Variables	200
11.5.7	Imputing Nominal and Ordinal Variables	203
11.5.8	Combining Imputed Values of Ordinal and Nominal Variables	203
11.6	Predicting Missing Values	204

## CHAPTER

## 12

## PREDICTIVE POWER AND VARIABLE REDUCTION I 207

---

12.1	Introduction	207
12.2	Metrics of Predictive Power	208

12.3	Methods of Variable Reduction	209
12.4	Variable Reduction: Before or During Modeling	210

## CHAPTER 13

### **ANALYSIS OF NOMINAL AND ORDINAL VARIABLES** 211

13.1	Introduction	211
13.2	Contingency Tables	211
13.3	Notation and Definitions	212
13.4	Contingency Tables for Binary Variables	214
13.4.1	Difference in Proportion	215
13.4.2	The Odds Ratio	218
13.4.3	The Pearson Statistic	221
13.4.4	The Likelihood Ratio Statistic	224
13.5	Contingency Tables for Multicategory Variables	225
13.6	Analysis of Ordinal Variables	227
13.7	Implementation Scenarios	231

## CHAPTER 14

### **ANALYSIS OF CONTINUOUS VARIABLES** 233

14.1	Introduction	233
14.2	When Is Binning Necessary?	233
14.3	Measures of Association	234
14.3.1	Notation	234
14.3.2	The $F$ -Test	236
14.3.3	Gini and Entropy Variances	236
14.4	Correlation Coefficients	239

## CHAPTER 15

### **PRINCIPAL COMPONENT ANALYSIS** 247

15.1	Introduction	247
15.2	Mathematical Formulations	248

15.3	Implementing and Using PCA	249
15.4	Comments on Using PCA	254
15.4.1	Number of Principal Components	254
15.4.2	Success of PCA	254
15.4.3	Nominal Variables	256
15.4.4	Dataset Size and Performance	256

## CHAPTER

## 16

<b>FACTOR ANALYSIS</b>	<b>257</b>
------------------------	------------

---

16.1	Introduction	257
16.1.1	Basic Model	257
16.1.2	Factor Rotation	259
16.1.3	Estimation Methods	259
16.1.4	Variable Standardization	259
16.1.5	Illustrative Example	259
16.2	Relationship Between PCA and FA	263
16.3	Implementation of Factor Analysis	263
16.3.1	Obtaining the Factors	264
16.3.2	Factor Scores	265

## CHAPTER

## 17

<b>PREDICTIVE POWER AND VARIABLE REDUCTION II</b>	<b>267</b>
---	------------

---

17.1	Introduction	267
17.2	Data with Binary Dependent Variables	267
17.2.1	Notation	267
17.2.2	Nominal Independent Variables	268
17.2.3	Numeric Nominal Independent Variables	273
17.2.4	Ordinal Independent Variables	273
17.2.5	Continuous Independent Variables	274
17.3	Data with Continuous Dependent Variables	275
17.3.1	Nominal Independent Variables	275
17.3.2	Ordinal Independent Variables	275
17.3.3	Continuous Independent Variables	275
17.4	Variable Reduction Strategies	275

## CHAPTER

**18****PUTTING IT ALL TOGETHER**

279

---

18.1	Introduction	279
18.2	The Process of Data Preparation	279
18.3	Case Study: The Bookstore	281
18.3.1	The Business Problem	281
18.3.2	Project Tasks	282
18.3.3	The Data Preparation Code	283

**APPENDIX****LISTING OF SAS MACROS**

297

---

A.1	Copyright and Software License	297
A.2	Dependencies between Macros	298
A.3	Data Acquisition and Integration	299
A.3.1	Macro TBRollup()	299
A.3.2	Macro ABRollup()	301
A.3.3	Macro VarMode()	303
A.3.4	Macro MergeDS()	304
A.3.5	Macro ContcatDS()	304
A.4	Integrity Checks	304
A.4.1	Macro SchCompare()	304
A.4.2	Macro CatCompare()	306
A.4.3	Macro ChiSample()	307
A.4.4	Macro VarUnivar1()	308
A.4.5	Macro CVLimits()	309
A.4.6	Macro CompareTwo()	309
A.5	Exploratory Data Analysis	310
A.5.1	Macro Extremes1()	310
A.5.2	Macro Extremes2()	311
A.5.3	Macro RobReg0L()	312
A.5.4	Macro Clust0L()	312
A.6	Sampling and Partitioning	313
A.6.1	Macro RandomSample()	313
A.6.2	Macro R2Samples()	313
A.6.3	Macro B2Samples()	315
A.7	Data Transformations	318
A.7.1	Macro NorList()	318

A.7.2	Macro NorVars()	319
A.7.3	Macro AutoInter()	320
A.7.4	Macro CalcCats()	321
A.7.5	Macro MapCats()	322
A.7.6	Macro CalcLL()	323
A.7.7	Macro BoxCox()	324
A.8	Binning and Reduction of Cardinality	325
A.8.1	Macro GRedCats()	325
A.8.2	Macro GSplit()	329
A.8.3	Macro AppCatRed()	331
A.8.4	Macro BinEqW()	332
A.8.5	Macro BinEqW2()	332
A.8.6	Macro BinEqW3()	333
A.8.7	Macro BinEqH()	334
A.8.8	Macro GBinBDV()	336
A.8.9	Macro AppBins()	340
A.9	Treatment of Missing Values	341
A.9.1	Macro ModeCat()	341
A.9.2	Macro SubCat()	342
A.9.3	Macro SubCont()	342
A.9.4	Macro MissPatt()	344
A.9.5	Macro ReMissPat()	347
A.9.6	Macro CheckMono()	349
A.9.7	Macro MakeMono()	350
A.9.8	Macro ImpReg()	351
A.9.9	Macro AvgImp()	351
A.9.10	Macro NORDImp()	352
A.10	Analysis of Nominal and Ordinal Variables	352
A.10.1	Macro ContinMat()	352
A.10.2	Macro PropDiff()	353
A.10.3	Macro OddsRatio()	354
A.10.4	Macro PearChi()	355
A.10.5	Macro LikeRatio()	355
A.10.6	Macro ContPear()	356
A.10.7	Macro ContSpear()	356
A.10.8	Macro ContnAna()	357
A.11	Analysis of Continuous Variables	358
A.11.1	Macro ContGrF()	358
A.11.2	Macro VarCorr()	359

A.12 Principal Component Analysis	360
A.12.1 Macro PrinComp1()	360
A.12.2 Macro PrinComp2()	360
A.13 Factor Analysis	362
A.13.1 Macro Factor()	362
A.13.2 Macro FactScore()	362
A.13.3 Macro FactRen()	363
A.14 Predictive Power and Variable Reduction II	363
A.14.1 Macro GiniCatBDV()	363
A.14.2 Macro EntCatBDV()	364
A.14.3 Macro PearSpear()	366
A.14.4 Macro PowerCatBDV()	367
A.14.5 Macro PowerOrdBDV()	368
A.14.6 Macro PowerCatNBDV()	370
A.15 Other Macros	372
A.15.1 ListToCol()	372
<b>BIBLIOGRAPHY</b>	373
<b>INDEX</b>	375
<b>ABOUT THE AUTHOR</b>	393

# LIST OF FIGURES

2.1	Steps of data flow	10
3.1	Decision tree diagram	21
3.2	A single neuron	23
3.3	Multilayered perceptron network with a single output	24
4.1	Resolution of macro variables with double coding	38
5.1	Data rollup	48
8.1	Balanced sampling	105
8.2	Data consistency constraints	105
9.1	Effect of Box–Cox transformation	138
9.2	Effect of Log transformation	139
9.3	Effect of power transformation	140
10.1	A dataset split	145
10.2	Splitting algorithm	146
10.3	Details of binary DV splits	147
11.1	Multiple imputation strategy	183
11.2	Combining imputed values for a continuous variable	200
14.1	Correlation between continuous variables	240
14.2	Correlation between a continuous variable and a binary variable	242
14.3	Correlation between a continuous variable and a binary variable	243
15.1	Scree plot of the credit card data	255
15.2	Scree plot of uncorrelated data	255
18.1	Data preparation process	280



This Page Intentionally Left Blank

# LIST OF TABLES

5.1	A sample of banking transactions	48
5.2	Result of rolling up the data of Table 5.1	49
5.3	Parameters of macro TBRollup()	52
5.4	Result of rollup macro	54
5.5	Parameters of macro VarMode()	55
5.6	Two sample tables: Left and Right	57
5.7	Result of merging: dataset “Both”	58
5.8	Table: Top	59
5.9	Table: Bottom	59
5.10	Table: Both	60
6.1	Parameters of macro SchCompare()	67
6.2	Parameters of macro CatCompare()	71
6.3	Results of PROC FREQ on SampleA and SampleB datasets	74
6.4	SampleA and SampleB frequencies as a contingency table	74
6.5	Parameters of macro ChiSample()	75
6.6	Common measures of quantifying the distribution of a continuous variable	77
6.7	Comparison of measures of continuous variables	78
6.8	Parameters of macro CVLimits()	79
6.9	Parameters of macro Compare Two()	81
7.1	Parameters of macro Extremes1()	88
7.2	Parameters of macro RobRegOL()	93
7.3	Parameters of macro ClustOL()	94
8.1	Parameters of macro R2samples()	102
8.2	Parameters of macro B2samples()	106
8.3	Distribution of marital status categories	113
9.1	Seven income values	117
9.2	Normalized income values	117

9.3	A convention for naming new variables	120
9.4	Parameters of macro <code>NorList()</code>	121
9.5	Parameters of macro <code>NorVars()</code>	122
9.6	Parameters of macro <code>AutoInter()</code>	124
9.7	Mapping of residence types	127
9.8	Parameters of macro <code>CalcCats()</code>	128
9.9	Parameters of macros <code>MappCats()</code>	128
9.10	Parameters of macro <code>CalcLL()</code>	134
9.11	Parameters of macro <code>BoxCox()</code>	135
9.12	Useful Transformation	140
10.1	Notation of splits with binary DV	148
10.2	Parameters of macro <code>GRedCats()</code>	149
10.3	Parameters of macro <code>GSplit()</code>	153
10.4	Parameters of macro <code>AppCatRed()</code>	155
10.5	Parameters of macro <code>BinEqW2()</code>	158
10.6	Equal-width bins	160
10.7	Equal-height bins	160
10.8	Parameters of macro <code>BinEqH()</code>	162
10.9	Parameters of macro <code>GBinBDV()</code>	165
10.10	Parameters of macros <code>AppBins()</code>	170
11.1	Five variables with missing values	181
11.2	Reordered five variables with missing values	181
11.3	Imputation methods available in PROC MI	182
11.4	Parameters of macro <code>MissPatt()</code>	185
11.5	The dataset <code>MP</code> as calculated by macro <code>MissPatt()</code>	190
11.6	Parameters of macro <code>ReMissPat()</code>	190
11.7	The dataset <code>Re_MP</code> as calculated by macro <code>ReMissPat()</code>	194
11.8	Parameters of macro <code>CheckMono()</code>	195
11.9	Parameters of macro <code>AvgImp()</code>	201
12.1	Common predictive power metrics	209
13.1	Gender distribution of mailing campaign results	212
13.2	Contingency table notation	213
13.3	Parameters of the macro <code>ContinMat()</code>	214
13.4	Success rate for each gender	215
13.5	Parameters of macro <code>PropDiff()</code>	216
13.6	Gender distribution of mailing campaign results	218

13.7	Parameters of macro OddsRatio()	220
13.8	Response versus credit card type	221
13.9	Actual expected counts of response status versus credit card type	222
13.10	Parameters of meters of macro PearChi()	223
13.11	Parameters of macro LikeRatio()	225
13.12	Parameters of macro ContnAna()	226
13.13	Credit card usage rate versus credit default	228
13.14	Parameters of the macro ContPear()	229
13.15	Parameters of the macro ContSpear()	230
14.1	Notation for association measures of continuous variables	234
14.2	Parameters of the marco ContGrF()	236
14.3	Incomes and home values	241
14.4	Values of Table 14.3 with two changes	241
14.5	Parameters of macro VarCorr()	244
15.1	Data of 20 credit card customers	250
15.2	Parameters of macro PrinComp2()	253
16.1	Parameters of macro Factor()	264
16.2	Parameters of macro FactScore()	265
16.3	Parameters of macro FactRen()	266
17.1	Parameters of macro GiniCatBDV()	269
17.2	Parameters of macro EntCatBDV()	271
17.3	Parameters of macro PearSpear()	274
17.4	Parameter of macro PowerCatBDV()	276

This Page Intentionally Left Blank

# PREFACE

The use of data mining modeling in different areas of business has moved over the last two decades from the domain of a few academic researchers to that of a large number of business and quantitative analysts in commercial organizations. This move has been accelerated by the rapid development of reliable commercial data mining software suites that allow easy implementation of many data mining algorithms such as regression, decision trees, neural networks, and cluster analysis.

However, data mining analysts spend up to 80% of their time, not doing actual modeling, but *preparing data*. This is due to three main reasons. First, all current modeling methods assume the data is in the form of a matrix containing the analysis variables, which somehow have been cleaned and prepared for modeling from their original source. The second reason is the nature of data sources. Almost all business data today is stored in relational form, either in a database, in a data warehouse, or in the back end of operational systems. Therefore, collecting the data from several tables from the relational form and converting data into the final matrix form is an essential task. Finally, the quality of the resulting models always depends on the quality of the data. This is not a surprise or a deviation from the familiar warning: *Garbage in, garbage out!* So, it turns out that data mining is mostly about data preparation.

Therefore, increasing the efficiency as well as the quality of data preparation procedures is at least as, if not more, important as making sure that the modeling algorithms are used to their fullest potential. This is the focus of this book.

This book is intended as a book of recipes. It provides the explanation of each data preparation procedure, as well as the SAS implementations, in the form of a macro that is suitable for task automation. Much of the code presented in this book can also be implemented in other data manipulation tools such as SQL. However, I have chosen SAS because of its dominance in today's data mining environment.

Although the focus of the book is on developing good mining views for data mining modeling, the presented techniques could also be used for the purpose of the other complementary task of reporting. Indeed, one of the essential data mining tasks is the generation of reports, both before and after the development and deployment of models.

One final note: this is not a book to teach readers how to use SAS or to do macro programming. It is a book about data preparation for data mining; therefore, a basic knowledge of the SAS system is assumed.