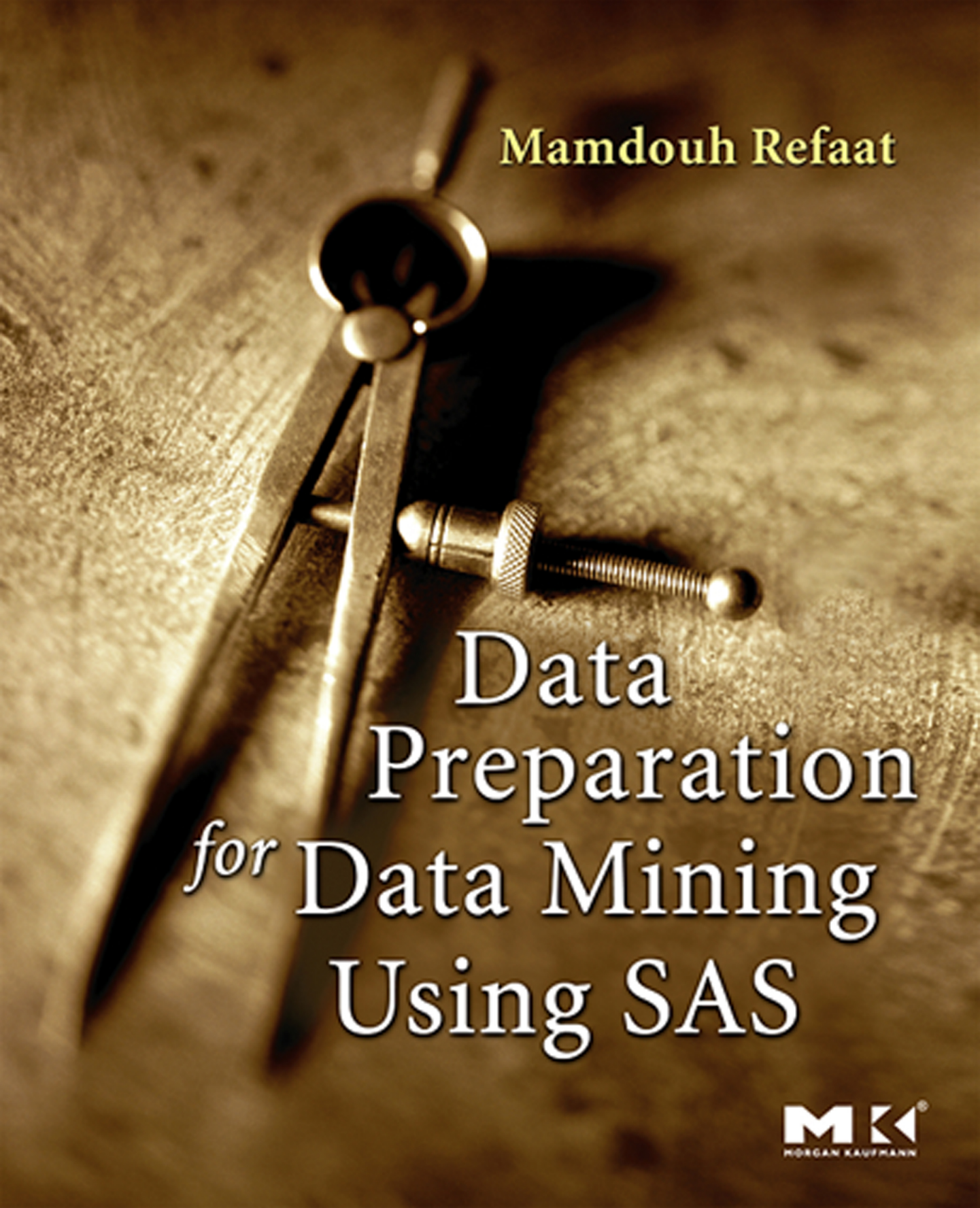
Mamdouh Refaat
Data Preparation for Data Mining Using SAS
MK
MORGAN KAUFMANN

# 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 Microsft® 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: IBMs 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 $O_2$*
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.



**Library of Congress Cataloging-in-Publication Data**
Refaat, Mamdouh.
Data preparation for data mining using SAS / Mamdouh Rafaat.
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* or *www.books.elsevier.com*

Printed in the United States of America
06 07 08 09 10 5 4 3 2 1

# Contents

# List of Figures

This Page Intentionally Left Blank

# List of Tables

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.