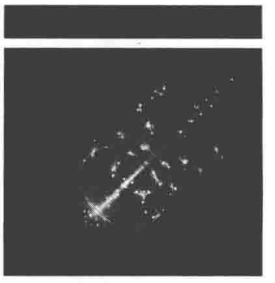
WILEY

TIMELY. PRACTICAL. RELIABLE.

Data Analysis Using SQL and Excel®

Second Edition

Gordon S. Linoff



Data Analysis Using SQL and Excel®

Second Edition

Gordon S. Linoff

WILEY

Data Analysis Using SQL and Excel®, Second Edition

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

ISBN: 978-1-119-02143-8

ISBN: 978-1-119-02145-2 (ebk)

ISBN: 978-1-119-02144-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2015950486

Data Analysis Using SQL and Excel®

To Giuseppe—for twenty five years, five books, and counting . . .

About the Author

Gordon S. Linoff has been working with databases, big data, and data mining for almost longer than he can remember. With decades of experience on the practice of using data effectively, he is a recognized expert in the field of data mining.

Gordon started using spreadsheets while a student at MIT, on the original Compaq Portable, the world's first luggable computer. Not very many years later, he managed a development group at the now-defunct Thinking Machines Corporation, tasked with building a massively parallel relational database for decision support.

After Thinking Machines' demise, he founded Data Miners in 1998 with his friend and former colleague Michael J. A. Berry (who left in 2012). Since then, he has worked on a wide diversity of projects across many different companies. He has taught hundreds of classes around the world on data mining and survival analysis through SAS Institute, a leader in statistical and business analytics software. He is also an avid contributor to Stack Overflow, particularly on questions related to databases, having the highest score in 2014.

Together with Michael Berry, Gordon has written several influential books on data mining, including *Data Mining Techniques for Marketing, Sales, and Customer Support*, the first book on data mining to achieve a third edition.

Gordon lives in New York with Giuseppe Scalia, his partner of 25 years.

Credits

Project Editor
John Sleeva

Technical Editor
Michael Berry

Production Editor
Dassi Zeidel

Copy Editor
Mike La Bonne

Manager of Content Development & Assembly
Mary Beth Wakefield

Marketing Director
David Mayhew

Marketing Manager
Carrie Sherrill

Professional Technology & Strategy Director
Barry Pruett

Business Manager
Amy Knies

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Brent Savage

Proofreader
Sara Wilson

Indexer
Johnna VanHoose Dinse

Cover Designer
Wiley

Cover Image
©iStock.com/Nobi_Prizue

Acknowledgments

Although this book has only one name on the cover, many people have helped me both specifically on this book and more generally in understanding data, analysis, and presentation.

I first met Michael Berry in 1990. We later founded Data Miners together, and he has been helpful on all fronts. He reviewed the chapters, tested the SQL code in the examples, and helped anonymize the data. His insights have been helpful and his debugging skills have made the examples much more accurate. His wife, Stephanie Jack, also deserves special praise for her patience and willingness to share Michael's time.

The original idea for the book came from Nick Drake, who then worked at Datran Media. A statistician by training, Nick was looking for a book that would help him use databases for data analysis. Bob Elliott, at the time my editor at Wiley, liked the idea.

Throughout the chapters, the understanding of data processing is based on dataflows, which Craig Stanfill of Ab Initio Corporation first introduced me to long ago when we worked together at Thinking Machines Corporation.

Along the way, I have learned a lot from many people. Anne Milley of SAS Institute first suggested that I learn survival analysis. Will Potts, now working at CapitalOne, then taught me much of what I know about the subject. Brij Masand helped extend the ideas to practical forecasting applications. Chi Kong Ho and his team at the *New York Times* provided valuable feedback for applying survival analysis to customer value calculations.

Stuart Ward from the *New York Times* and Zaiying Huang spent countless hours explaining and discussing statistical concepts. Harrison Sohmer, also of the *New York Times*, taught me many Excel tricks, some of which I've been able to include in the book.

Jamie MacLennan and the SQL Server team at Microsoft have been helpful in answering my questions about the product.

Over the past few years, I have been a major contributor to Stack Overflow. Along the way, I have learned an incredible amount about SQL and about how to explain concepts. A handful of people whom I've never met in person have helped in various ways. Richard Stallman invented emacs and the Free Software Foundation; emacs provided the basis for the calendar table. Rob Bovey of Applications Professional, Inc. created the X-Y chart labeler used in several chapters. The Census data set was created by the folks at the Missouri Census Data Center. Juice Analytics inspired the example for Worksheet bar charts in Chapter 5 (and thanks to Alex Wimbush, who pointed me in their direction). Edwin Straver of Frontline Systems answered several questions about Solver.

Over the years, many colleagues, friends, and students have provided inspiration, questions, and answers. There are too many to list them all, but I want to particularly thank Eran Abikhzer, Christian Albright, Michael Benigno, Emily Cohen, Carol D'Andrea, Sonia Dubin, Lounette Dyer, Victor Fu, Josh Goff, Richard Greenburg, Gregory Lampshire, Mikhail Levdanski, Savvas Mavridis, Fiona McNeill, Karen Kennedy McConlogue, Steven Mullaney, Courage Noko, Laura Palmer, Alan Parker, Ashit Patel, Ronnie Rowton, Vishal Santoshi, Adam Schwebber, Kent Taylor, John Trustman, John Wallace, David Wang, and Zhilang Zhao. I would also like to thank the folks in the SAS Institute Training group who have organized, reviewed, and sponsored my data mining classes for many years, giving me the opportunity to meet many interesting and diverse people involved with data mining.

I also thank all those friends and family I've visited while writing this book and who (for the most part) allowed me the space and time to work—my mother, my father, my sister Debbie, my brother Joe, my in-laws Raimonda Scalia, Ugo Scalia, and Terry Sparacio, and my friends Jon Mosley, Paul Houlihan, Leonid Poretsky, Anthony DiCarlo, and Maciej Zworski. On the other hand, my cat Luna, who spent many hours curled up next to me, will miss my writing.

Finally, acknowledgments would be incomplete without thanking Giuseppe Scalia, my partner through seven books, who has managed to maintain my sanity through all of them.

Thank you, everyone!

Foreword

Gordon Linoff and I have written three and a half books together. (Four, if we get to count the second edition of *Data Mining Techniques* as a whole new book; it didn't feel like any less work.) Neither of us has written a book without the other before, so I must admit to a tiny twinge of regret upon first seeing the cover of this one without my name on it next to Gordon's. The feeling passed very quickly as recollections of the authorial life came flooding back—vacations spent at the keyboard instead of in or on the lake, opportunities missed, relationships strained. More importantly, this is a book that only Gordon Linoff could have written. His unique combination of talents and experiences informs every chapter.

I first met Gordon at Thinking Machines Corporation, a now long-defunct manufacturer of parallel supercomputers where we both worked in the late eighties and early nineties. Among other roles, Gordon managed the implementation of a parallel relational database designed to support complex analytical queries on very large databases. The design point for this database was radically different from other relational database systems available at the time in that no trade-offs were made to support transaction processing. The requirements for a system designed to quickly retrieve or update a single record are quite different from the requirements for a system to scan and join huge tables. Jettisoning the requirement to support transaction processing made for a cleaner, more efficient database for analytical processing. This part of Gordon's background means he understands SQL for data analysis literally from the inside out.

Just as a database designed to answer big important questions has a different structure from one designed to process many individual transactions, a *book* about using databases to answer big important questions requires a different approach to SQL. Many books on SQL are written for database administrators.

Others are written for users wishing to prepare simple reports. Still others attempt to introduce some particular dialect of SQL in every detail. This one is written for data analysts, data miners, and anyone who wants to extract maximum information value from large corporate databases. Jettisoning the requirement to address all the disparate types of database users makes this a better, more focused book for the intended audience. In short, this is a book about how to use databases the way we ourselves use them.

Even more important than Gordon's database technology background are his many years experience as a data mining consultant. This has given him a deep understanding of the kinds of questions businesses need to ask and of the data they are likely to have available to answer them. Years spent exploring corporate databases have given Gordon an intuitive feel for how to approach the kinds of problems that crop up time and again across many different business domains:

- **How to take advantage of geographic data.** A zip code field looks much richer when you realize that from zip code you can get to latitude and longitude, and from latitude and longitude you can get to distance. It looks richer still when your realize that you can use it to join in Census Bureau data to get at important attributes, such as population density, median income, percentage of people on public assistance, and the like.
- **How to take advantage of dates.** Order dates, ship dates, enrollment dates, birth dates. Corporate data is full of dates. These fields look richer when you understand how to turn dates into tenures, analyze purchases by day of week, and track trends in fulfillment time. They look richer still when you know how to use this data to analyze time-to-event problems such as time to next purchase or expected remaining lifetime of a customer relationship.
- **How to build data mining models directly in SQL.** This book shows you how to do things in SQL that you probably never imagined possible, including generating association rules for market basket analysis, building regression models, and implementing naïve Bayesian models and scorecards.
- **How to prepare data for use with data mining tools.** Although more than most people realize can be done using just SQL and Excel, eventually you will want to use more specialized data mining tools. These tools need data in a specific format known as a *customer signature*. This book shows you how to create these data mining extracts.

The book is rich in examples and they all use real data. This point is worth saying more about. Unrealistic datasets lead to unrealistic results. This is frustrating to the student. In real life, the more you know about the business context, the better your data mining results will be. Subject matter expertise gives you a head start. You know what variables ought to be predictive and have good ideas

about new ones to derive. Fake data does not reward these good ideas because patterns that should be in the data are missing and patterns that shouldn't be there have been introduced inadvertently. Real data is hard to come by, not least because real data may reveal more than its owners are willing to share about their business operations. As a result, many books and courses make do with artificially constructed datasets. Best of all, the datasets used in the book are all available for download at `www.wiley.com/go/dataanalysisusingsqlandexcel2e`.

I reviewed the chapters of this book as they were written. This process was very beneficial to my own use of SQL and Excel. The exercise of thinking about the fairly complex queries used in the examples greatly increased my understanding of how SQL actually works. As a result, I have lost my fear of nested queries, multi-way joins, giant case statements, and other formerly daunting aspects of the language. In well over a decade of collaboration, I have always turned to Gordon for help using SQL and Excel to best advantage. Now, I can turn to this book. And you can, too.

—Michael J. A. Berry