Gerald V. Post
Data Mining Applications/2e
Analyzing Business Data

# Data Mining Applications

## Analyzing Business Data

**Version 2.0.0**

Gerald V. Post

*University of the Pacific*

Data Mining Applications/2e
Analyzing Business Data

Copyright © 2013 by Gerald V. Post
All rights reserved. No part of this publication may be reproduced or distributed in any form or stored in any database or retrieval system without the prior written consent of Gerald V. Post.

Students:
Your honesty is critical to your reputation. No company wants to hire a thief—particularly for jobs as critical as application development and database administration. If someone is willing to steal something as inexpensive as an e-book, how can that person be trusted with billions of dollars in corporate accounts?

You are not allowed to "share" this book in any form with anyone else. You cannot give or sell any information from this publication in any form to anyone else.

A note on the cover photo:

Even in a wildflower preserve you can find patterns and clusters.

Or you can think of data mining as a way to search through acres of hills, flowers, sand, and ocean to find interesting facts.

Or can data mining software find the physical site of the image?

To purchase this book or other books: http://JerryPost.com/Books

For my students.

# Preface



## Business Applications of Data Mining

Most managers commonly report that they are buried in data. Database management systems, enterprise resource planning systems, and Web sites now collect and store transactions from all aspects of business and organizations. Managers have access to terabytes (or more) of historical data. But how is this data supposed to be used? Database management systems can create standard reports and charts, and query languages (SQL) can be used to answer specific questions. But it is not enough.

Managers need statistics to analyze data—but few managers are statisticians, and configuring and interpreting traditional statistical tools can require extensive experience. Data mining (DM) tools were created partly from statistics and partly from computer science to analyze large data sets. Some tools can run relatively unsupervised—simply select the data and run the routines to obtain results and recommendations. Other tools require more supervision, from selecting variables to tweaking models to interpreting results and reconfiguring the analysis.

If you want a job as a manager, you need to know how to use information technology. But it is not simply a matter of knowing how to use a word processor or spreadsheet. You need to use the technology to collaborate with other workers, to analyze data, and to find ways to improve your organization.

Over the past decade, the field of data mining, or business intelligence or analytics, has received increased attention by academics, managers, and tool vendors. Notably, the tools have improved dramatically in terms of ease of use and interpretation. Many books and Web sites focus on data mining; but most of them emphasize either the statistics or the computer science issues. Few of them deal with the managerial issues of how to configure the tools, apply them to business problems, and interpret the results to make decisions. This focus is the main purpose of this book.

This book primarily focuses on using SQL Server Analysis Services—the data mining component of SQL Server. This tool is useful because it is widely available to universities through the MSDN Academic Alliance. Trial versions are also available to anyone via the Microsoft Web site. It is also relatively easy to use—although a few tips and tricks are necessary to fully understand some of the tools. Other tools, including the open-source Weka tool are used to demonstrate a few techniques. Many other tools exist and could be used to analyze the data provided with this book.

## Learning Assessment

After finishing the book, students should be able to understand the primary data mining tools, know what types of problems they are used to evaluate, and understand how to interpret the output and apply the results of the tools to help make business decisions.

Anyone who teaches this material knows that this learning objective is difficult to meet—because of the huge number of possible issues and the flexibility required in analyzing problems. To meet the objective, smaller, and more concrete goals are presented in each chapter. Each of these is spelled out as a series of questions at the start of the chapter. By the end of each chapter, students should be able to provide intelligent answers to the various questions.

# About the Book

## Organization

The book begins with an introduction to database queries because queries are often used to configure data for specific tools and because queries are useful for finding answers to specific questions. Students do not need a database course to be able to use this book. But, this book only lightly covers queries and just touches on database design issues. Chapter 3 explains the purpose and configuration of OLAP data cubes and the basic cube browsers. It emphasizes the importance of identifying facts and dimensions in any problem and then shows how the cube browser computes subtotals and filters data. It also explains how to configure hierarchies of various data dimensions. Chapter 4 reviews basic concepts in probability and statistics. The objective is to explain all of the concepts needed in the rest of the book. It would be best if students have already had a separate course in probability and statistics. .


Chapter 1: Introduction

Part One: Fundamental Tools
Chapter 2 Finding and Storing Data
Chapter 3 OLAP Cubes
Chapter 4 Probability and Statistics Summary

Part Two: Business Analysis
Chapter 5 Cluster Analysis
Chapter 6 Association and Market Baskets
Chapter 7 Evaluation of Dimensions
Chapter 8 Time Series Analysis

Part Three: Specialized Tools
Chapter 9 Geographic and Sequence Analysis
Chapter 10 Multidimensional Expressions (MDX)


The second section explores the fundamental data mining tools used in business analysis. Note that the chapters are organized by business goals as opposed to by individual tools. Chapter 5 examines basic clustering tools used to define groups—it is particularly useful in marketing to identify groups of customers that have certain attributes in common. Chapter 6 covers association or market basket analysis. It explains how probability computations are used to identify which items are associated or commonly purchased together. Chapter 7 is a long chapter because it explores how dimensions are related to fact dimensions using basic tools such as regression, logistic regression, and neural networks for nonlinear analysis. Chapter 8 covers the basics of time series analysis.

Clearly, all tools covered in this section could be explained in greater detail. In fact, individual books have been written about each type of analysis. But, the point of this book is not to make managers experts in every single tool. Instead, the goal is to make it possible for managers to apply the right technique needed for a specific decision and to correctly interpret the results and make better decisions.

Part Three is new in the second edition. Chapter 9 covers geographic information. This section is more about visualization and display of data than about searching. Many items in business are related to location and it is often helpful to examine decisions in that context. The chapter also introduces sequence analysis. Much of the hard-core work in sequences has been done in biology—specifically DNA analysis. But some of the tools are useful for certain business problems such as tracing Web browsing patterns. Chapter 10 introduces the basic foundations of multidimensional expressions (MDX). This language was initially defined by

Microsoft but has been adopted by many vendors. It is useful for creating complex computations that are difficult to evaluate with other methods..

## Features That Focus on Solving Problems

Why do students need a textbook? Most definitions exist on the Web and many Web sites and blogs contain information on data mining, individual tools, and even how to solve specific problems. And yes, if you are serious about data mining, you should use all of these resources to learn as much as possible about the details. But students still need a textbook. This textbook provides a structure to the information. It teaches you how the many pieces fit together to solve business problems. It provides the context and ideas on how to analyze the millions of tidbits of data floating on the Web. Use the textbook to provide the foundation knowledge that describes data mining, how to configure data, and how to interpret results. As technology evolves and new problems are encountered, use the Internet to share ideas and solutions with others.

Each chapter contains several unique features to assist in understanding the material and in applying it to analyze data and make business decisions. The most important parts of each chapter are the examples. The chapters explain the basic problem, how to configure the data, how to run the analysis, and how to interpret the results. The writing explains the process, but it generally avoids a step-by-step presentation. The problem with step-by-step presentations is that readers focus too heavily on tracking steps and losing attention on the overall process. Also, as tools change, the steps quickly become obsolete, although the overall process remains the same.

Each chapter contains various features to explain the purpose of the chapter and help students understand and review the main concepts. The basic features are:

- **What you will learn in this chapter**. A series of questions highlight the important issues.
- **Chapter summary**. A brief synopsis of the chapter highlights—useful when reviewing for exams.
- **Key Words**. A list of words introduced in the chapter. A full glossary is provided at the end of the text.
- **Review Questions**. Designed as a study guide of the main topics in the chapter.
- **Exercises**. Problems that apply the knowledge learned in the chapter. Each set of exercises begins with a list of exercises that mirror the actions taken in that chapter (Book Exercises). Additional exercises using the main cases are next, followed by a set of exercises designed to be solved by a team of students. The team exercises involve work that can be partitioned among team members and combined to achieve a final answer.
- **Additional Reading**. References for more detailed investigation of the topics.

## This Book Is Different from Other Texts

First, this book is a business text designed for an upper division or MBA course. It is not a statistics book and it is not a computer science book. Those types of books are useful and some excellent versions exist, but they are less useful to business students.

Second, this book relies on commonly available tools along with several realistic data sets to illustrate how to analyze business data. Students can use tools that already exist, apply them to the data sets provided, and immediately begin analyzing the results. Many other books require students to obtain tools with limited licenses and restricted capabilities, or to write their own programs. The goal is to get students analyzing business data as quickly and easily as possible, using commonly available tools.

### Instructor Resources

Instructors have access to the following resources for course presentation and management. All the instructor supplements were created by the author, except the test bank:

- Instructor's Manual includes answers to all end-of-chapter review questions, exercises, and teaching notes for the industry-specific cases. Teaching tips and ties to the PowerPoint slides are included for each chapter.
- A test bank contains true/false, multiple choice, and short answer questions.
- Lecture notes are available as slide shows in Microsoft PowerPoint format. The slides contain all of the figures along with additional notes. The slides are organized into lectures and can be rearranged to suit individual preferences.
- Several databases and exercises are available online. The instructor can add new data, modify the exercises, or use them to expand on the discussion in the text. Most of the primary databases were created with custom data generators. These data generators are available to instructors (typically within Microsoft Access databases) so that new data with new patterns can be generated.
- The book's Web site at JerryPost.com provides resources for instructors and students using the text.

## Installing the Sample Databases

The book uses several databases to demonstrate how to set up various tools and how to interpret results. These databases are used throughout the book and as exercises in each chapter. The individual databases are described in Chapter 1. Some of the databases were chosen to emphasize specific data mining topics. The Rolling Thunder Bicycle company database is used more than the others because it contains a wider variety of data. The data will be easier to understand if you learn a little about bicycles. Some background information is on the Web site: http://www.JerryPost.com.

In every case, you should approach the situation as if you were a manager or consultant. If necessary, visit a similar organization, talk to a manager or employee in the field, read annual reports for a firm in the industry, or find additional background information on the Web. It is easier to understand and apply the results when you understand the goals and decisions facing the organization.

### Software

The majority of the book relies on Microsoft SQL Server and its associated SQL Server Analysis Services (SSAS). When SSAS tools depart from traditional applications, the book also uses traditional techniques to show the difference. The tools chosen, including SQL Server, were specifically selected because they are readily available to students and they are relatively easy to use.

*Microsoft SQL Server*

SQL Server can be obtained for educational use through the MSDN Academic Alliance, now known as DreamSpark. This organization does require a small fee ($500 first year, $320 a year for renewal), but it provides a license for all of Microsoft's development tools, including the right to provide copies to students. Many schools are already members, but additional details can be found at: https://www.dreamspark.com.

If you are working alone or unable to join MSDN-AA, a trial version of SQL Server can be downloaded free from Microsoft. It is valid for 180 days. Note that the SQL Server Express version will not work. You need the Standard, Enterprise, or Developer version. The Standard version does not support all of the Analysis Services options, but only a couple of them are used in this book. The examples in this book were built with the SQL Server 2012 Developer's Edition, which is equivalent to the Enterprise edition.

Installing SQL Server on a single computer is relatively straightforward. Just be sure to install the Analysis Services. Also, it is useful to specify Mixed Mode security—which sets SQL Server to accept logins using both Windows accounts and internal SQL Server accounts. When installing Reporting Services and Analysis Services, be sure to specify at least one administrator account that you control. The best approach is to install everything onto a single computer and give yourself administrator rights to each tool. This approach reduces the learning hurdles by minimizing interference from the security systems.

In a class environment, it is possible to install a single copy of the SQL Server database engine, Reporting Services, and Analysis Services on a central server. Individual student computers can then run just the Visual Studio and SQL Server Management Studio components and attach directly to the central server. However, it is necessary to configure security on the server to give relatively open access to each student. Do not use a production server! For the purposes of learning the tools, it is generally simpler to install all of the components on each student's computer and giving the student full control over that instance.

The first time you create an Analysis Services project usually requires some additional configuration steps. Some of the steps require experimentation. The most difficult one is the connection from the Analysis Services system to the database (impersonation). The tool provides four choices. You might have to test all four choices before finding one that works for your installation. The examples in the book suggest one choice that often works, but it does not work in all cases.

*Weka*

Weka is an interesting open source tool available from the University of Waikato in New Zealand at http://www.cs.waikato.ac.nz/ml/weka/. It is available for several operating systems and uses the GNU general public license. It supports several standard data mining algorithms for classification, clustering, association, and attribute selection. It also has some useful graphics tools for visualization. It can read some specialized data files, but is easiest to use with flat CSV files. Installation is straightforward—simply download and run the installation package from the Web site.

*gretl*

Data mining is heavily based on statistics and econometrics has developed several tools to analyze data. The details of many of the tools and the problems they

solve are beyond the scope of this book. However, many of the tools are relatively automated and can be used to handle specialized problems. High-end (read "expensive") statistical packages have been used for these tasks. However, Allin Cottrell and Riccardo Lucchetti have created the open source tool gretl that is both free and easy to use. It is available at http://gretl.soruceforge.net. Tasks involving regression and time series analysis are particularly useful in gretl. Again, the tool is easiest to use with flat CSV files. The process of extracting the data from SQL Server and creating a CSV file is explained in this book when the tool is used.

### *R*

R or R System is not used in this book. However, it is a powerful open source statistics package that many other people are using for data mining. R is available free from http://www.r-project.org. The GUI interface Rattle is also useful and it is available from http://rattle.togaware.com. R is often run on the Linux platform, but versions exist for Windows. The interesting aspect of R is that it is designed for people to create new tools and algorithms. By itself, it is largely a platform. Developers create new tools that can be added to the system to analyze data. This approach makes it easy for others to contribute to the development of new tools, so R can grow. But, it makes it harder for new students to select the proper tool.

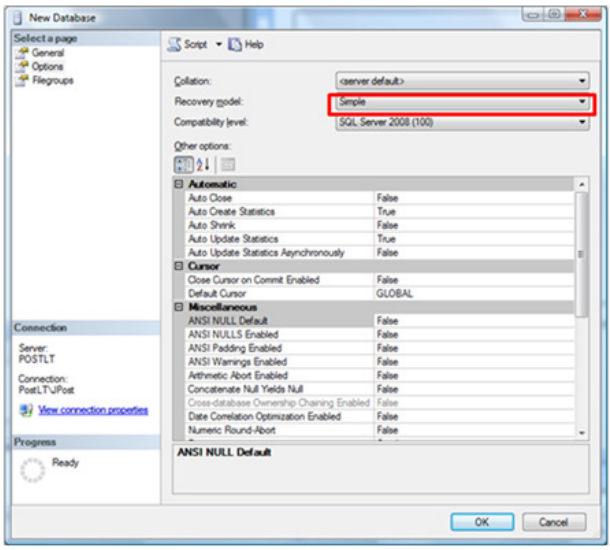
## Database Installation

The data for the databases and set up scripts are available for download from the book's Web site. Each database is packed into a compressed ZIP file. Download each file and extract its contents into a separate folder. The contents consist of several CSV files and a setup SQL script. The script creates the individual tables in SQL Server and bulk loads the data from the CSV files. However, you have to make a couple of changes to the script to tell it where you stored the files. The process is described here for a single database. Repeat it for each database you wish to use.

Start SQL Server Management Studio. Log in with an account that has at least Create Database permissions—generally you want to use a DBA account, or the Windows account that was used to install SQL Server. Create a new database—right-click the Databases entry and choose New Database. Enter a name that represents the database, such as RT for Rolling Thunder. If you are building the database on a central server, you might want to change the file locations to improve performance, but it is not critical.

A more important change is to set the Recovery Mode. The default mode in SQL Server is designed for transaction databases where all data changes are logged to protect the data in the event of failures. This approach creates log files that can be several times larger than the database table files. As with most data warehouses, you can always reload the data, so recovery is unimportant. As shown in the figure, click the Options page and change the Recovery mode to Simple (instead of Full). This mode does not write any data to the log file and saves a considerable amount of space—particularly for the Bakery database because it has several million rows of data.

After the database has been created, you can load the data. Create a new query. The easiest method is to right-click the new database and choose New Query. Load the SQL script from the download into the query window. One quick method is to use the Windows Explorer to navigate to the downloaded files and drag the SQL file into the query window. You need to change two things in the script that tailor




it to your situation: <DB>, the name of the database, and <PATH>, the location of the CSV files.

Use Ctrl-H (or Edit/Find and Replace/Quick Replace) to start the replacement tool. Enter <DB> as the search text—including the brackets. Enter the name of your database (such as RT or Bakery). Click the Replace All button. Repeat the process and search for <PATH> with the brackets, and enter the full local path name of the folder that holds the CSV files (such as C:\Downloads\Bakery\). You can usually copy the path name from Windows Explorer, but you need to add the trailing back slash (\). Click the button to replace all instances. Note that the CSV files should be on the same computer as the SQL Server program. If you are using a central server, you should copy the CSV folder to that physical server. Otherwise, you need to specify a network path and assign appropriate security permissions. When <DB> and <PATH> have been replaced, execute the script. After a few minutes, you will see a list of the tables and the number of rows. If you get error messages, read them to check for obvious problems. Then reload the script, and run the two replacement commands more carefully. The scripts have been tested on many different machines. Problems that arise are usually due to errors in the replacement path or to local security issues.

# Brief Contents

1 Introduction

## Part One: Fundamental Tools

2 Database Management Systems

3 OLAP Cubes

4 Probability and Statistics Summary

## Part Two: Business Analysis

5 Cluster Analysis

6 Association and Market Baskets

7 Evaluation of Dimensions

8 Time Series Analysis

## Part Three: Specialized Tools

9 Sequence and Geographic Analysis

10 Multidimensional Expressions (MDX)

# Contents

Business Applications of Data Mining, v
Learning Assessment, v
About the Book, vi
*Organization, vi*
*Features That Focus on Solving Problems, vii*
*This Book Is Different from Other Texts, vii*
*Instructor Resources, viii*
Installing the Sample Databases, viii
*Software, viii*
*Database Installation , x*

**Introduction, 1**

Introduction, 3
Business Situation, 4
*Finance, Risk, and Costs, 4*
*Marketing, 6*
*Production and Supply Chain Management, 11*
*Human Resources Management, 12*
Perspectives, 14
*Probability and Statistics, 15*
*Machine Learning, 16*
*Computer Science: Challenges of Large Data Sets, 16*
*Management Applications, 16*
Database, 17
*Traditional Transactions Processing, 18*
*Data Warehouse and Analytical Processing, 19*
*Data Sources, 19*
*Data Extraction, Transformation, and Loading, 19*
Software Tools, 20
*Data Mining Techniques, 20*
*Data Mining Tools, 21*
*Statistical Tools, 22*
*Production Systems and Scaling, 22*
Potential Dangers, 24
*Human Errors , 24*
*Insufficient Data, 25*
*Bad Data , 25*
*Over Fitting, 26*
*Random Chance, 27*
*Estimation Instability, 28*
*Model Instability, 29*
Introduction to Cases, 30
*Rolling Thunder Bicycle Company, 30*
*Diner, 31*
*Corner Med, 31*
*Basketball, 32*
*Bakery, 32*
*Cars, 32*
Summary, 33
Key Words, 34
Review Questions, 34
Exercises, 34
Additional Reading, 37

**Fundamental Tools, 39**

**Database Management Systems, 40**

Introduction, 42
Relational Databases, 43
*Tables, 43*
*Data Types, 45*
Four Questions to Retrieve Data, 47
*What Output Do You Want to See?, 48*
*What Do You Already Know?, 48*
*What Tables Are Involved?, 48*
*How Are the Tables Joined?, 48*
Query Basics, 49
*Single Tables, 50*
*Introduction to SQL, 52*
*Sorting the Output, 52*
*Criteria, 53*
*Useful WHERE Clauses, 56*
Computations, 57
*Basic Arithmetic Operators, 57*
*Aggregation, 58*
*Functions, 60*
Subtotals and GROUP BY, 61
*Conditions on Totals (HAVING), 63*
*WHERE versus HAVING, 64*
*The Best and the Worst, 65*
Multiple Tables, 66
*Joining Tables, 66*
*Identifying Columns in Different Tables, 67*
*Joining Many Tables, 68*
*Views: Saved Queries, 69*
*LEFT JOIN, 71*
*UNION, 71*

Data Manipulation, 72
*UPDATE, 73*
*INSERT, 73*
*DELETE, 74*
SQL Server Reports, 75
*Administration Configuration, 75*
*Creating a Report, 76*
Database Design Concepts, 82
*Notation, 84*
*First Normal Form, 86*
*Second Normal Form, 86*
*Third Normal Form, 87*
Summary, 88
Key Words, 89
Review Questions, 90
Exercises, 90
Additional Reading, 93

**OLAP Cubes, 94**

Introduction, 96
Challenges with the Relational Model, 98
*Indexes, 99*
*Data Warehouse, 99*
*Extraction, Transformation, and Loading, 100*
*MOLAP, ROLAP, and HOLAP, 101*
OLAP Design, 102
*Facts and Dimensions, 104*
*Star Design, 105*
*Snowflake Design, 107*
*Hierarchies, 107*
Creating a Cube with Microsoft Analysis Services, 108
*Data Sources, 109*
*Data Source Views, 111*
*Cubes, 116*
Dimensions, 120
*Hierarchies, 123*
*Time Dimensions, 124*
*Custom Geographic Hierarchy, 128*
*Attribute Relationships, 130*
Fine Tuning the Cube, 133
*Calculations and Queries, 134*
*Perspectives, 138*
*Internationalization and Translations, 140*
*Performance: Partitions and Aggregations, 142*
Excel PivotTables, 144
Actions, 146
Key Performance Indicators, 149
*Definition, 149*
*Creating KPIs, 150*
*Browsing a KPI, 153*
Summary, 154
Key Words, 155
Review Questions, 155
Exercises, 156
Additional Reading, 158

**Probability and Statistics Summary, 159**

Introduction, 161
Probability Basics, 161
*Discrete and Continuous Data, 163*
*Counting and Combinations, 163*
*Probability Rules, 166*
Interdependencies: Joint Probabilities, 169
*Contingency Tables, 170*
*Tree Diagrams, 171*
*Bayes Theorem, 172*
Probability Distributions, 176
*Discrete Data, 176*
*Continuous Data, 181*
*Joint and Conditional Probabilities, 182*
*Expected Value (Mean) and Variance, 183*
*Important Continuous Distributions, 187*
Statistics, 195
*Samples, 195*
*Common Statistics, 196*
*Confidence Intervals, 198*
*Hypothesis Testing, 200*
*Chi-Square Hypothesis Tests, 203*
*Information Measure, 205*
Summary, 206
Key Words, 208
Review Questions, 208
Exercises, 209
Additional Reading, 214

**Business Analysis, 215**

**Cluster Analysis, 216**

Introduction, 218
Business Situation, 220

Model, 221
*Distance or Dissimilarities, 222*
*Combinatorial Searches with K-Means, 224*
*Statistical Mixture Model with EM, 227*
*Hierarchical Clusters, 229*
*Other Statistical Methods, 233*
Data, 236
*Attributes and Observations, 236*
*Continuous and Discrete Data, 237*
*Missing Data, 238*
Clustering on Products: Cars, 238
*Goals, 238*
*Data, 239*
*Microsoft Clustering, 241*
*Results from Microsoft Clustering, 243*
*Prediction, 246*
*Larger Model and Parameter Changes, 248*
Traditional EM Clustering, 251
*Goals and Data, 252*
*Results, 254*
*K-Means Clusters, 255*
Comparison, 256
Customer Clustering with Categorical Data, 258
*Data, 258*
*Microsoft Clustering Results, 259*
*Weka Clustering Results, 260*
Summary, 262
Key Words, 263
Review Questions, 263
Exercises, 264
Additional Reading, 267

**Association and Market Baskets, 268**

Introduction, 270
Business Situation, 271
*The Bakery, 272*
*Product and Dimension Levels, 273*
Model, 274
*Goal, 275*
*Assigning Values to Rules, 276*
Problems with Dimensions, 282
*The A Priori Algorithm, 283*
*Issues in Setting Minimum Support and Confidence, 284*
Potential Problems, 285
*Simpson's Paradox, 285*
*Skewed Support Data, 286*
*Continuous Data, 288*
*Quantity, 289*
Data, 290
*Database Structure, 291*
*Market Basket Structure, 291*
Traditional Tools for Association Rules, 292
*Goals, 293*
*Data, 293*
*Results, 293*
Microsoft Association Rules, 294
*Goals, 295*
*Data, 295*
*Results, 296*
*Comparing Results, 298*
Summary, 300
Key Words, 301
Review Questions, 301
Exercises, 302
Additional Reading, 305

**Evaluation of Dimensions, 306**

Introduction, 308
Business Situation, 309
Model, 309
Data, 311
*Attributes and Observations, 312*
*Continuous and Discrete Data, 312*
*Missing Data, 313*
Linear Regression, 313
*Goals, 314*
*Data, 316*
*Tools, 318*
*Results, 322*
*Attribute Evaluation, 327*
*Prediction, 328*
Logistic Regression, 330
*Goals, 330*
*Data, 332*
*Tools, 333*
*Results, 334*
*Attribute Evaluation, 338*
*Prediction, 339*
Naïve Bayes, 340
*Goals, 341*
*Data, 345*

*Tools, 345*
*Results, 346*
*Attribute Evaluation, 346*
*Prediction, 348*
Decision Trees, 349
*Goals, 350*
*Data, 352*
*Tools, 353*
*Results, 353*
*Attribute Evaluation, 355*
*Prediction, 355*
Neural Network, 358
*Goals, 359*
*Data, 361*
*Tools, 361*
*Results, 361*
*Attribute Evaluation, 362*
*Prediction, 363*
Model Comparisons, 365
*Prediction, 366*
*Attribute Evaluation, 367*
*Nonlinear Complications, 368*
Summary, 369
Key Words, 370
Review Questions, 370
Exercises, 371
Additional Reading, 373

**Time Series Analysis, 375**

Introduction, 377
Business Situation, 378
Model, 379
*Time Series Components, 379*
*Auto Regression, 381*
*Moving Average, 384*
*Trends, 387*
*ARIMA, 389*
*Cross Correlations, 392*
*Evaluating Models, 395*
Data, 396
*Attributes and Observations, 396*
*Missing Data, 397*
Traditional ARIMA Estimation, 397
*Goals, 398*
*Tools, 398*
*Results, 400*
*Forecasts, 404*
*Seasonality Evaluation, 405*
Microsoft Time Series Estimation, 406
*Goals, 406*
*Data, 408*
*Tools, 409*
*Results, 412*
*ARTxp Model, 413*
*Forecasts, 415*
*Seasonality Evaluation, 417*
Cross Correlation and Linear Regression, 417
*Goals, 418*
*Data, 418*
*Tools, 419*
Comparison, 425
Summary, 426
Key Words, 427
Review Questions, 428
Exercises, 428
Additional Reading, 431

**Specialized Tools, 432**

**Sequence Analysis and GIS, 433**

Introduction, 435
Sequence Analysis , 435
Business Situation, 437
Model, 438
*Classification, 439*
*Clustering: Business examples, 440*
Data, 442
*Attributes and Observations, 442*
*Continuous and Discrete Data, 444*
*Missing Data, 444*
*Web Log Data Files, 445*
SQL Server Sequence Clustering, 447
*Goals, 448*
*Data, 448*
*Tools, 449*
*Results, 452*
*Prediction, 457*
Other Tools, 458
Sequence Summary, 458
Geographic Analysis, 459
Business Situation, 461
Data, 461

Model, 462
Microsoft MapPoint, 464
Other Tools, 467
*Esri, 468*
*Google, 471*
*Bing, 472*
*Federal Government, 473*
Geographic Summary, 474
Key Words, 474
Review Questions, 475
Exercises, 476
Additional Reading, 479

**MDX, 480**

Introduction, 482
OLAP Cube Review, 482
*Dimensions and Hierarchies, 484*
*Rolling Thunder Bicycles Cube, 484*
Definitions and Concepts, 487
Main Syntax, 489
Basic Examples, 490
*A First Example, 491*
*Adding a WHERE Condition, 492*
*Displaying Specific Dimension Values, 494*
*Cross Join, 494*
Calculated Measures, 495
Complex Computations, 496
*Percentages, 497*
*Compute Changes, 498*
*ParallelPeriod Function, 499*
Some MDX Functions, 501
*EXCEPT: Taking Values Out of Totals, 501*
*Conditions with IIF and CoalesceEmpty, 502*
*TopCount Function, 504*
*Year to Date, 505*
*Moving Averages, 507*
Summary, 509
Key Words, 510
Review Questions, 511
Exercises, 512
Additional Reading, 515