## Contents

### Foreword

xxxiii

### Introduction

xxxvii

### Chapter 1: A Data Miner Looks at SQL

1

Databases, SQL, and Big Data

2

What Is Big Data?

3

Relational Databases

3

Hadoop and Hive

4

NoSQL and Other Types of Databases

4

SQL

5

Picturing the Structure of the Data

6

What Is a Data Model?

6

What Is a Table?

7

Allowing NULL Values

8

Column Types

9

What Is an Entity-Relationship Diagram?

10

The Zip Code Tables

12

Subscription Dataset

13

Purchases Dataset

14

Tips on Naming Things

14

Picturing Data Analysis Using Dataflows

16

What Is a Dataflow?

16

READ: Reading a Database Table

18

OUTPUT: Outputting a Table (or Chart)

18

SELECT: Selecting Various Columns in the Table

18

FILTER: Filtering Rows Based on a Condition

18

APPEND: Appending New Calculated Columns

19

UNION: Combining Multiple Datasets into One

19

AGGREGATE: Aggregating Values

19
LOOKUP: Looking Up Values in One Table in Another 19
CROSSJOIN: Generating the Cartesian Product of Two Tables 19
JOIN: Combining Two Tables Using a Key Column 20
SORT: Ordering the Results of a Dataset 20
Dataflows, SQL, and Relational Algebra 20
SQL Queries 21
What to Do, Not How to Do It 21
The SELECT Statement 22
A Basic SQL Query 22
A Basic Summary SQL Query 24
What It Means to Join Tables 25
Cross-Joins: The Most General Joins 26
Lookup: A Useful Join 27
Equijoins 29
Nonequijoins 31
Outer Joins 31
Other Important Capabilities in SQL 32
UNION ALL 33
CASE 33
IN 34
Window Functions 35
Subqueries and Common Table Expressions 36
Are Our Friends 36
Subqueries for Naming Variables 37
Subqueries for Handling Summaries 40
Subqueries and IN 42
Rewriting the “IN” as a JOIN 42
Correlated Subqueries 43
NOT IN Operator 44
EXISTS and NOT EXISTS Operators 45
Subqueries for UNION ALL 46
Lessons Learned 47

Chapter 2 What’s in a Table? Getting Started with Data Exploration 49
What Is Data Exploration? 50
Excel for Charting 51
A Basic Chart: Column Charts 51
Inserting the Data 52
Creating the Column Chart 53
Formatting the Column Chart 55
Bar Charts in Cells 57
Character-Based Bar Charts 57
Conditional Formatting-Based Bar Charts 58
Useful Variations on the Column Chart 59
A New Query 59
Side-by-Side Columns
Contents

- Stacked Columns 60
- Stacked and Normalized Columns 60
- Number of Orders and Revenue 60
- Other Types of Charts 63
  - Line Charts 63
  - Area Charts 63
  - X-Y Charts (Scatter Plots) 64
- Sparklines 65
- What Values Are in the Columns? 68
  - Histograms 68
  - Histograms of Counts 72
  - Cumulative Histograms of Counts 74
  - Histograms (Frequencies) for Numeric Values 75
    - Ranges Based on the Number of Digits, Using Numeric Techniques 75
    - Ranges Based on the Number of Digits, Using String Techniques 77
    - More Refined Ranges: First Digit Plus Number of Digits 77
    - Breaking Numeric Values into Equal-Sized Groups 77
- More Values to Explore—Min, Max, and Mode 79
  - Minimum and Maximum Values 79
  - The Most Common Value (Mode) 80
    - Calculating Mode Using Basic SQL 80
    - Calculating Mode Using Window Functions 81
- Exploring String Values 81
  - Histogram of Length 82
  - Strings Starting or Ending with Spaces 82
  - Handling Upper- and Lowercase 82
  - What Characters Are in a String? 83
- Exploring Values in Two Columns 86
  - What Are Average Sales by State? 86
  - How Often Are Products Repeated within a Single Order? 86
    - Direct Counting Approach 87
    - Comparison of Distinct Counts to Overall Counts 88
  - Which State Has the Most American Express Users? 89
- From Summarizing One Column to Summarizing All Columns 90
  - Good Summary for One Column 90
  - Query to Get All Columns in a Table 93
  - Using SQL to Generate Summary Code 94
- Lessons Learned 96

Chapter 3 How Different Is Different? 97
- Basic Statistical Concepts 98
  - The Null Hypothesis 98
  - Confidence and Probability 100
  - Normal Distribution 101
How Different Are the Averages? 105  
   The Approach 105  
   Standard Deviation for Subset Averages 105  
   Three Approaches 107  
      Estimation Based on Two Samples 108  
      Estimation Based on Difference 109  
Sampling from a Table 110  
   Random Sample 110  
   Repeatable Random Sample 111  
   Proportional Stratified Sample 112  
   Balanced Sample 113  
Counting Possibilities 115  
   How Many Men? 116  
   How Many Californians? 120  
   Null Hypothesis and Confidence 122  
   How Many Customers Are Still Active? 123  
      Given the Count, What Is the Probability? 124  
      Given the Probability, What Is the Number of Stops? 125  
   The Rate or the Number? 126  
Ratios and Their Statistics 128  
   Standard Error of a Proportion 128  
   Confidence Interval on Proportions 129  
   Difference of Proportions 131  
   Conservative Lower Bounds 132  
Chi-Square 132  
   Expected Values 133  
   Chi-Square Calculation 134  
   Chi-Square Distribution 134  
   Chi-Square in SQL 135  
   What States Have Unusual Affinities for Which Types of Products? 138  
      Data Investigation 138  
      SQL to Calculate Chi-Square Values 139  
   Affinity Results 140  
What Months and Payment Types Have Unusual Affinities for Which Types of Products? 140  
   Multidimensional Chi-Square 141  
   Using a SQL Query 141  
   The Results 142  
Lessons Learned 143  

Chapter  4  Where Is It All Happening? Location, Location, Location 145  
Latitude and Longitude 146  
   Definition of Latitude and Longitude 146  
   Degrees, Minutes, Seconds, and All That 147  
   Distance between Two Locations 149
Contents

Euclidian Method 149
Accurate Method 151
Finding All Zip Codes within a Given Distance 152
Finding Nearest Zip Code in Excel 154
Pictures with Zip Codes 155
The Scatter Plot Map 155
Who Uses Solar Power for Heating? 157
Where Are the Customers? 159
Census Demographics 160
The Extremes: Richest and Poorest 161
Median Income 161
Proportion of Wealthy and Poor 162
Income Similarity and Dissimilarity Using Chi-Square 163
Comparison of Zip Codes with and without Orders 167
Zip Codes Not in Census File 167
Profiles of Zip Codes with and without Orders 168
Classifying and Comparing Zip Codes 170
Geographic Hierarchies 172
Wealthiest Zip Code in a State? 172
Zip Code with the Most Orders in Each State 175
Interesting Hierarchies in Geographic Data 176
Counties 177
Designated Marketing Areas 177
Census Hierarchies 178
Other Geographic Subdivisions 178
Geography on the Web 179
Calculating County Wealth 181
Identifying Counties 181
Measuring Wealth 182
Distribution of Values of Wealth 183
Which Zip Code Is Wealthiest Relative to Its County? 185
County with Highest Relative Order Penetration 185
Mapping in Excel 188
Why Create Maps? 188
It Can't Be Mapped 190
Mapping on the Web 190
State Boundaries on Scatter Plots of Zip Codes 191
Plotting State Boundaries 191
Pictures of State Boundaries 193
Lessons Learned 194

Chapter 5  It's a Matter of Time 197
Dates and Times in Databases 198
Some Fundamentals of Dates and Times in Databases 199
Extracting Components of Dates and Times 199
Converting to Standard Formats 201
## Contents

Intervals (Durations) 202  
Time Zones 203  
Calendar Table 203  

### Starting to Investigate Dates 204  
Verifying That Dates Have No Times 204  
Comparing Counts by Date 205  
Order Lines Shipped and Billed 206  
Customers Shipped and Billed 208  
Number of Different Bill and Ship Dates per Order 209  

### Counts of Orders and Order Sizes 210  
Items as Measured by Number of Units 211  
Items as Measured by Distinct Products 211  
Size as Measured by Dollars 214  

### Days of the Week 215  
Billing Date by Day of the Week 215  
Changes in Day of the Week by Year 216  
Comparison of Days of the Week for Two Dates 217  

### How Long Between Two Dates? 218  
Duration in Days 218  
Duration in Weeks 220  
Duration in Months 221  
How Many Mondays? 221  
A Business Problem about Days of the Week 222  
Outline of a Solution 222  
Solving It in SQL 224  
Using a Calendar Table Instead 224  
When Is the Next Anniversary (or Birthday)? 225  
First Year Anniversary This Month 225  
First Year Anniversary Next Month 226  
Manipulating Dates to Calculate the Next Anniversary 227  

### Year-over-Year Comparisons 229  
Comparisons by Day 229  
Adding a Moving Average Trend Line 230  
Comparisons by Week 231  
Comparisons by Month 231  
Month-to-Date Comparison 233  
Extrapolation by Days in Month 235  
Estimation Based on Day of Week 237  
Estimation Based on Previous Year 239  

### Counting Active Customers by Day 239  
How Many Customers on a Given Day? 239  
How Many Customers Every Day? 240  
How Many Customers of Different Types? 241  
How Many Customers by Tenure Segment? 242  
Calculating Actives Entirely Using SQL 246  

Simple Chart Animation in Excel 247
Contents

Order Date to Ship Date 248
Order Date to Ship Date by Year 250
Querying the Data 250
Creating the One-Year Excel Table 251
Creating and Customizing the Chart 252

Lessons Learned 254

Chapter 6 How Long Will Customers Last? Survival Analysis to Understand Customers and Their Value 255
Background on Survival Analysis 256
  Life Expectancy 256
  Medical Research 258
  Examples of Hazards 259
The Hazard Calculation 260
  Data Investigation 261
    Stop Flag 261
    Tenure 262
  Hazard Probability 264
  Visualizing Customers: Time versus Tenure 265
  Censoring 266
Survival and Retention 269
  Point Estimate for Survival 269
  Calculating Survival for All Tenures 269
  Calculating Survival in SQL 271
    Calculating the Product of Column Values 272
    Adding in More Dimensions 274
  A Simple Customer Retention Calculation 274
Comparison between Retention and Survival 276
Simple Example of Hazard and Survival 276
  Constant Hazard 277
  What Happens to a Mixture? 278
    Constant Hazard Corresponding to Survival 279
Comparing Different Groups of Customers 280
  Summarizing the Markets 280
  Stratifying by Market 281
  Survival Ratio 284
    Conditional Survival 285
Comparing Survival over Time 287
  How Has a Particular Hazard Changed over Time? 288
  What Is Customer Survival by Year of Start? 289
  What Did Survival Look Like in the Past? 290
Important Measures Derived from Survival 293
  Point Estimate of Survival 293
  Median Customer Tenure 294
  Average Customer Lifetime 295
  Confidence in the Hazards 297
Using Survival for Customer Value Calculations
  Estimated Revenue
  Estimating Future Revenue for One Future Start
    Value in the First Year
    SQL Day-by-Day Approach
  Estimated Revenue for a Group of Existing Customers
    Estimated Second Year Revenue for a Homogenous Group
  Estimated Future Revenue for All Customers
Forecasting
  Existing Base Forecast
    Existing Base Calculation
    Calculating Survival on July 1st
    Calculating the Number of Existing Customers on July 1st
  How Good Is It?
    Estimating the Long-Term Hazard
  New Start Forecast
Lessons Learned

Chapter 7  Factors Affecting Survival: The What and Why of Customer Tenure
Which Factors Are Important and When
  Explanation of the Approach
  Using Averages to Compare Numeric Variables
    The Answer
    Answering the Question in SQL and Excel
    Answering the Question Entirely in SQL
    Extension to Include Confidence Bounds
  Hazard Ratios
    Interpreting Hazard Ratios
    Calculating Hazard Ratios Using SQL and Excel
    Calculating Hazard Ratios in SQL
    Why the Hazard Ratio?
Left Truncation
  Recognizing Left Truncation
  Effect of Left Truncation
  How to Fix Left Truncation, Conceptually
Estimating Hazard Probability for One Tenure
Estimating Hazard Probabilities for All Tenures
Doing the Calculation in SQL
Time Windowing
  A Business Problem
  Time Windows = Left Truncation + Right Censoring
    Calculating One Hazard Probability Using a Time Window
    All Hazard Probabilities for a Time Window
    Comparison of Hazards by Stops in Year in Excel
    Comparison of Hazards by Stops in Year in SQL
### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competing Risks</td>
<td>342</td>
</tr>
<tr>
<td>Examples of Competing Risks</td>
<td>342</td>
</tr>
<tr>
<td>I=Involuntary Churn</td>
<td>343</td>
</tr>
<tr>
<td>V=Voluntary Churn</td>
<td>343</td>
</tr>
<tr>
<td>M=Migration</td>
<td>344</td>
</tr>
<tr>
<td>Other</td>
<td>344</td>
</tr>
<tr>
<td>Competing Risk “Hazard Probability”</td>
<td>345</td>
</tr>
<tr>
<td>Competing Risk “Survival”</td>
<td>346</td>
</tr>
<tr>
<td>What Happens to Customers over Time</td>
<td>347</td>
</tr>
<tr>
<td>Example</td>
<td>347</td>
</tr>
<tr>
<td>A Cohort-Based Approach</td>
<td>348</td>
</tr>
<tr>
<td>The Survival Analysis Approach</td>
<td>351</td>
</tr>
<tr>
<td>Before and After</td>
<td>353</td>
</tr>
<tr>
<td>Three Scenarios has</td>
<td>353</td>
</tr>
<tr>
<td>A Billing Mistake</td>
<td>353</td>
</tr>
<tr>
<td>A Loyalty Program</td>
<td>354</td>
</tr>
<tr>
<td>Raising Prices</td>
<td>355</td>
</tr>
<tr>
<td>Using Survival Forecasts to Understand One-Time Events</td>
<td>356</td>
</tr>
<tr>
<td>Forecasting Identified Customers Who Stopped</td>
<td>356</td>
</tr>
<tr>
<td>Estimating Excess Stops</td>
<td>357</td>
</tr>
<tr>
<td>Before and After Comparison</td>
<td>357</td>
</tr>
<tr>
<td>Cohort-Based Approach</td>
<td>358</td>
</tr>
<tr>
<td>Cohort-Based Approach: Full Cohorts</td>
<td>358</td>
</tr>
<tr>
<td>Direct Estimation of Event Effect</td>
<td>361</td>
</tr>
<tr>
<td>Approach to the Calculation</td>
<td>361</td>
</tr>
<tr>
<td>Time-Dependent Covariate Survival Using SQL and Excel</td>
<td>362</td>
</tr>
<tr>
<td>Doing the Calculation in SQL</td>
<td>364</td>
</tr>
<tr>
<td>Lessons Learned</td>
<td>366</td>
</tr>
</tbody>
</table>

#### Chapter 8  Customer Purchases and Other Repeated Events  367

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifying Customers</td>
<td>368</td>
</tr>
<tr>
<td>Who Is the Customer?</td>
<td>368</td>
</tr>
<tr>
<td>How Many?</td>
<td>369</td>
</tr>
<tr>
<td>How Many Genders in a Household?</td>
<td>371</td>
</tr>
<tr>
<td>Investigating First Names</td>
<td>374</td>
</tr>
<tr>
<td>Other Customer Information</td>
<td>378</td>
</tr>
<tr>
<td>First and Last Names</td>
<td>378</td>
</tr>
<tr>
<td>Addresses</td>
<td>380</td>
</tr>
<tr>
<td>Email Addresses</td>
<td>381</td>
</tr>
<tr>
<td>Other Identifying Information</td>
<td>382</td>
</tr>
<tr>
<td>How Many New Customers Appear Each Year?</td>
<td>383</td>
</tr>
<tr>
<td>Counting Customers</td>
<td>383</td>
</tr>
<tr>
<td>Span of Time Making Purchases</td>
<td>386</td>
</tr>
<tr>
<td>Average Time between Orders</td>
<td>388</td>
</tr>
<tr>
<td>Purchase Intervals</td>
<td>390</td>
</tr>
<tr>
<td>How Many Days in a Row Do Customers Make Purchases?</td>
<td>391</td>
</tr>
</tbody>
</table>
RFM Analysis 393
   The Dimensions 394
      Recency 394
      Frequency 396
      Monetary 397
   Calculating the RFM Cell 398
   How Is RFM Useful? 399
      A Methodology for Marketing Experiments 400
      Customer Migration 400
      RFM Limits 403
Which Households Are Increasing Purchase Amounts Over Time? 404
   Comparison of Earliest and Latest Values 404
      Calculating the Earliest and Latest Values 404
      Comparing the First and Last Values 409
   Comparison of First Year Values and Last Year Values 413
   Trend from the Best Fit Line 415
      Using the Slope 415
      Calculating the Slope 415
Time to Next Event 416
   Idea behind the Calculation 417
      Calculating Next Purchase Date Using SQL 417
   From Next Purchase Date to Time-to-Event 418
   Stratifying Time-to-Event 419
Lessons Learned 420

**Chapter 9 What’s in a Shopping Cart? Market Basket Analysis** 421
Exploring the Products 422
   Scatter Plot of Products 422
   Which Product Groups Are Shipped in Which Years? 423
Duplicate Products in Orders 426
   Are Duplicates Explained by the Product? 427
   Are Duplicates Explained by the Product Group? 428
   Are Duplicates Explained by Timing? 429
   Are Duplicates Explained by Multiple Ship Dates or Prices? 430
Histogram of Number of Units 431
   Which Products Tend to be Sold Multiple Times Within an Order? 433
Changes in Price 435
Products and Customer Worth 437
   Consistency of Order Size 437
   Products Associated with One-Time Customers 440
   Products Associated with the Best Customer 444
   Residual Value 445
Product Geographic Distribution 448
   Most Common Product by State 448
Which Products Have Broad Appeal Versus Local Appeal 449
Which Customers Have Particular Products? 451
Which Customers Have the Most Popular Products? 451
Which Products Does a Customer Have? 453
Lists Using Conditional Aggregation 454
Aggregate String Concatenation in SQL Server 455
Lists Using String Aggregation in SQL Server 456
Which Customers Have Three Particular Products? 458
Three Products Using Joins 458
Three Products Using Exists 460
Using Conditional Aggregation and Filtering 461
Generalized Set-Within-a-Set Queries 462
Lessons Learned 463

Chapter 10 Association Rules and Beyond 465
Item Sets 466
- Combinations of Two Products 466
  - Number of Two-Way Combinations 466
  - Generating All Two-Way Combinations 467
  - Examples of Item Sets 469
- More General Item Sets 470
  - Combinations of Product Groups 470
  - Larger Item Sets 471
  - All Item Sets Up to a Given Size 474
- Households Not Orders 476
  - Combinations within a Household 476
  - Investigating Products within Households but Not within Orders 476
  - Multiple Purchases of the Same Product 478
- The Simplest Association Rules 480
  - Associations and Rules 480
  - Zero-Way Association Rules 481
  - What Is the Distribution of Probabilities? 481
  - What Do Zero-Way Associations Tell Us? 483
- One-Way Association Rules 483
  - Evaluating a One-Way Association Rule 483
  - Generating All One-Way Rules 485
  - One-Way Rules with Evaluation Information 486
  - One-Way Rules on Product Groups 488
- Two-Way Associations 489
  - Calculating Two-Way Associations 489
  - Using Chi-Square to Find the Best Rules 491
    - Applying Chi-Square to Rules 491
    - Applying Chi-Square to Rules in SQL 492
    - Comparing Chi-Square Rules to Lift 493
    - Chi-Square for Negative Rules 495
Heterogeneous Associations 496
   Rules of the Form “State Plus Product” 496
   Rules Mixing Different Types of Products 498
Extending Association Rules 499
   Multi-Way Associations 500
   Multi-Way Associations in One Query 501
   Rules Using Attributes of Products 502
   Rules with Different Left- and Right-Hand Sides 502
   Before and After: Sequential Associations 503
Lessons Learned 506

Chapter 11 Data Mining Models in SQL 507
Introduction to Directed Data Mining 508
   Directed Models 509
   The Data in Modeling 509
      Model Set 509
      Score Set 511
      Prediction Model Sets versus Profiling Model Sets 511
   Examples of Modeling Tasks 513
      Similarity Models 513
      Yes-or-No Models (Binary Response Classification) 513
      Yes-or-No Models with Propensity Scores 514
      Multiple Categories 514
      Estimating Numeric Values 515
   Model Evaluation 515
Look-Alike Models 515
   What Is the Model? 516
   What Is the Best Zip Code? 516
   A Basic Look-Alike Model 518
   Look-Alike Using Z-Scores 519
   Example of Nearest-Neighbor Model 521
Lookup Model for Most Popular Product 522
   Most Popular Product 522
   Calculating Most Popular Product Group 523
   Evaluating the Lookup Model 525
   Using a Profiling Lookup Model for Prediction 525
   Using Binary Classification Instead 526
Lookup Model for Order Size 528
   Most Basic Example: No Dimensions 528
   Adding One Dimension 529
   Adding More Dimensions 531
   Examining Nonstationarity 531
   Evaluating the Model Using an Average Value Chart 532
Lookup Model for Probability of Response 534
   The Overall Probability as a Model 534
   Exploring Different Dimensions 535
How Accurate Are the Models? 537
ROC Charts and AUC 540
Creating an ROC Chart 540
Calculating Area under the Curve (AUC) 542
Adding More Dimensions 544
Naïve Bayesian Models (Evidence Models) 546
Some Ideas in Probability 546
Probabilities and Conditional Probabilities 546
Odds 548
Likelihood 548
Calculating the Naïve Bayesian Model 549
An Intriguing Observation 549
Bayesian Model of One Variable 551
Bayesian Model of One Variable in SQL 551
The “Naïve” Generalization 553
Naïve Bayesian Model: Scoring and Lift 555
Scoring with More Attributes 555
Creating a Cumulative Gains Chart 557
Comparison of Naïve Bayesian and Lookup Models 558
Lessons Learned 559

Chapter 12  The Best-Fit Line: Linear Regression Models 561
The Best-Fit Line 562
Tenure and Amount Paid 562
Properties of the Best-fit Line 563
What Does Best-Fit Mean? 563
Formula for Line 565
Expected Value 565
Error (Residuals) 567
Preserving Averages 568
Inverse Model 568
Beware of the Data 569
Trend Lines in Charts 571
Best-Fit Line in Scatter Plots 571
Logarithmic, Power, and Exponential Trend Curves 572
Polynomial Trend Curves 573
Moving Average 575
Best-Fit Using the LINEST() Function 577
Returning Values in Multiple Cells 577
Calculating Expected Values 578
LINEST() for Logarithmic, Exponential, and Power Curves 580
Measuring Goodness of Fit Using R² 581
The R² Value 581
Limitations of R² 582
What R² Really Means 583
Direct Calculation of Best-Fit Line Coefficients 584
Calculating the Coefficients 584
Calculating the Best-Fit Line in SQL 586
Price Elasticity 587
  Price Frequency 587
  Price Frequency for $20 Books 589
  Price Elasticity Model in SQL 591
  Price Elasticity Average Value Chart 591
Weighted Linear Regression 592
  Customer Stops during the First Year 593
Weighted Best Fit 594
  Weighted Best-Fit Line in a Chart 596
  Weighted Best-Fit in SQL 596
  Weighted Best-Fit Using Solver 597
    The Weighted Best-Fit Line 598
    Solver Is More Accurate Than a Guessing Game 599
More Than One Input Variable 600
  Multiple Regression in Excel 601
    Getting the Data 601
    Investigating Each Variable Separately 601
    Building a Model with Three Input Variables 603
    Using Solver for Multiple Regression 604
    Choosing Input Variables One-By-One 604
    Multiple Regression in SQL 605
Lessons Learned 607

Chapter 13 Building Customer
Signatures for Further Analysis 609
What Is a Customer Signature? 610
What Is a Customer? 611
  Sources of Data for the Customer Signature 612
    Current Customer Snapshot 612
    Initial Customer Information 613
    Self-Reported Information 614
    External Data (Demographic and So On) 614
    About Their Neighbors 615
    Transaction Summaries and Behavioral Data 615
Using Customer Signatures 616
  Data Mining Modeling 616
  Scoring Models 616
  Ad Hoc Analysis 616
    Repository of Customer-Centric Business Metrics 616
Designing Customer Signatures 617
  Profiling versus Prediction 617
  Column Roles 617
    Identification Columns 618
    Input Columns 618
Target Columns 618
Foreign Key Columns 618
Cutoff Date 618
Time Frames 619
  Naming of Columns 619
  Eliminating Seasonality 619
  Adding Seasonality Back In 620
  Multiple Time Frames 621
Operations to Build Customer Signatures 622
  Driving Table 622
    Using an Existing Table as the Driving Table 623
    Derived Table as the Driving Table 624
Looking Up Data 625
  Fixed Lookup Tables 625
  Customer Dimension Lookup Tables 627
Initial Transaction 628
Pivoting 629
  Payment Type Pivot 630
  Channel Pivot 632
  Year Pivot 633
  Order Line Information Pivot 634
Summarizing 637
  Basic Summaries 637
  More Complex Summaries 637
Extracting Features 639
  Geographic Location Information 640
Date Time Columns 640
Patterns in Strings 641
  Email Addresses 641
  Addresses 642
  Product Descriptions 642
  Credit Card Numbers 643
Summarizing Customer Behaviors 644
  Calculating Slope for Time Series 644
    Calculating Slope from Pivoted Time Series 644
    Calculating Slope for a Regular Time Series 647
    Calculating Slope for an Irregular Time Series 648
  Weekend Shoppers 648
  Declining Usage Behavior 650
Lessons Learned 653

Chapter 14  Performance Is the Issue: Using SQL Effectively 655
Query Engines and Performance 656
  Order Notation for Understanding Performance 656
A Simple Example 657
  Full Table Scan 658
Parallel Full Table Scan 658
Index Lookup 659
Performance of the Query 660
Considerations When Thinking About Performance 660
Storage Management (Memory and Disk) 660
Indexes 662
Processing Engine and Parallel Processing 663
Performance: Its Meaning and Measurement 663
Performance Improvement 101 665
Ensure that Types are Consistent 666
Reference Only the Columns and Tables That Are Needed by the Query 666
Use DISTINCT Only When Necessary 666
UNION ALL: 1, UNION: 0 667
Put Conditions in WHERE Rather Than HAVING 667
Use OUTER JOINs Only When Needed 667
Using Indexes Effectively 668
What Are Indexes? 668
B-Trees 668
Hash Indexes 670
Spatial Indexes (R-Trees) 670
Full Text Indexes (Inverted Indexes) 671
Variations on B-Tree Indexes 672
Simple Examples of Indexes 673
Equality in a Where Clause 673
Variations on the Theme of Equality 674
Inequality in a WHERE Clause 675
ORDER BY 675
Aggregation 676
Limitations on Indexes 676
Effectively Using Composite Indexes 679
Composite Indexes for a Query with One Table 679
Composite Indexes for a Query with Joins 681
When OR Is a Bad Thing 683
Sometimes UNION ALL Is Better Than OR 683
Sometimes LEFT OUTER JOIN Is Better Than OR 684
Sometimes Multiple Conditional Expressions Are Better 686
Pros and Cons: Different Ways of Expressing the Same Thing 686
What States Are Not Recognized in Orders? 687
The Most Obvious Query 687
A Simple Modification 687
A Better Version 687
An Alternative Using LEFT JOIN 688
How Big Would That Intermediate Table Be? 689
A GROUP BY Conundrum 689
A Basic Query 689
<table>
<thead>
<tr>
<th>Appendix</th>
<th>Equivalent Constructs Among Databases</th>
<th>703</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td></td>
<td>731</td>
</tr>
</tbody>
</table>