Contents

Preface xiii

Part 1 Data Preparation: Business Point of View

Chapter 1 Analytic Business Questions 3

1.1 Introduction 3

1.2 The Term *Business Question* 4

1.3 Examples of Analytic Business Questions 4

1.4 The Analysis Process 5

1.5 Challenging an Analytic Business Question 6

1.6 Business Point of View Needed 9

Chapter 2 Characteristics of Analytic Business Questions 11

2.1 Introduction 12

2.2 Analysis Complexity: Real Analytic or Reporting? 12

2.3 Analysis Paradigm: Statistics or Data Mining? 13

2.4 Data Preparation Paradigm: As Much Data As Possible or Business Knowledge First? 14

2.5 Analysis Method: Supervised or Unsupervised? 15

2.6 Scoring Needed: Yes/No? 16

2.7 Periodicity of Analysis: One-Shot Analysis or Re-run Analysis? 17

2.8 Need for Historic Data: Yes/No? 18

2.9 Data Structure: One-Row-per-Subject or Multiple-Rows-per-Subject? 19

2.10 Complexity of the Analysis Team 19

2.11 Conclusion 19

Chapter 3 Characteristics of Data Sources 21

3.1 Introduction 21

3.2 Operational or Dispositive Data Systems? 22

3.3 Data Requirement: Periodic Availability 24

3.4 Wording: Analysis Table or Analytic Data Mart? 25

3.5 Quality of Data Sources for Analytics 25

Chapter 4 Different Points of View on Analytic Data Preparation 29

4.1 Introduction 29

4.2 Simon, Daniele and Elias: Three Different Roles in the Analysis Process 30

4.3 Simon—The Business Analyst 30

4.4 Daniele—The Quantitative Expert 31

4.5 Elias—The IT and Data Expert 32

4.6 Who Is Right? 33

4.7 The Optimal Triangle 35

Part 2 Data Structures and Data Modeling

Chapter 5 The Origin of Data 39

5.1 Introduction 39

5.2 Data Origin from a Technical Point of View 40

5.3 Application Layer and Data Layer 40

5.4 Simple Text Files or Spreadsheets 40

5.5 Relational Database Systems 41

5.6 Enterprise Resource Planning Systems 41

5.7 Hierarchical Databases 42

5.8 Large Text Files 42

5.9 Where Should Data Be Accessed From? 43

Chapter 6 Data Models 45

6.1 Introduction 45

6.2 Relational Model and Entity Relationship Diagrams 46

6.3 Logical versus Physical Data Model 46

6.4 Star Schema 47

6.5 Normalization and De-normalization 49

Chapter 7 Analysis Subjects and Multiple Observations 51

7.1 Introduction 51

7.2 Analysis Subject 52

7.3 Multiple Observations 53

7.4 Data Mart Structures 55

7.5 No Analysis Subject Available? 59

Chapter 8 The One Row-per-Subject Data Mart 61

8.1 Introduction 61

8.2 The One-Row-per-Subject Paradigm 62

8.3 The Technical Point of View 64

8.4 The Business Point of View: Transposing or Aggregating Original Data 65

8.5 Hierarchies: Aggregating Up and Copying Down 67

8.6 Conclusion 68

Chapter 9 The Multiple-Rows-per-Subject Data Mart 69

9.1 Introduction 69

9.2 Using Multiple-Rows-per-Subject Data Marts 70

9.3 Types of Multiple-Rows-per-Subject Data Marts 71

9.4 Multiple Observations per Time Period 74

9.5 Relationship to Other Data Mart Structures 75

Chapter 10 Data Structures for Longitudinal Analysis 77

10.1 Introduction 77

10.2 Data Relationships in Longitudinal Cases 79

10.3 Transactional Data, Finest Granularity, and Most Appropriate Aggregation Level 82

10.4 Data Mart Structures for Longitudinal Data Marts 83

Chapter 11 Considerations for Data Marts 89

11.1 Introduction 89

11.2 Types and Roles of Variables in a Data Mart 89

11.3 Derived Variables 92

11.4 Variable Criteria 93

Chapter 12 Considerations for Predictive Modeling 95

12.1 Introduction 95

12.2 Target Windows and Observation Windows 96

12.3 Multiple Target Windows 97

12.4 Overfitting 98

Part 3 Data Mart Coding and Content

Chapter 13 Accessing Data 105

13.1 Introduction 105
13.2 Accessing Data from Relational Databases Using SAS/ACCESS Modules 106
13.3 Accessing Data from Microsoft Office 108
13.4 Accessing Data from Text Files 108
13.5 Accessing Data from Hierarchical Text Files 110
13.6 Other Access Methods 113

Chapter 14 Transposing One- and Multiple-Rows-per-Subject Data Structures 115

14.1 Introduction 115
14.2 Transposing from a Multiple-Rows-per-Subject Data Set to a One-Row-per-Subject Data Set 117
14.3 Transposing from a One-Row-per-Subject Data Set to a Multiple-Rows-per-Subject Data Set 120
14.4 Transposing a Transactional Table with Categorical Entries 124
14.5 Creating Key-Value Tables 128

Chapter 15 Transposing Longitudinal Data 131

15.1 Introduction 131
15.2 Standard Scenarios 132
15.3 Complex Scenarios 134

Chapter 16 Transformations of Interval-Scaled Variables 139

16.1 Introduction 139
16.2 Simple Derived Variables 140
16.3 Derived Relative Variables 143
16.4 Time Intervals 146
16.5 Binning Observations into Groups 149
16.6 Transformations of Distributions 153
16.7 Replacing Missing Values 159
16.8 Conclusion 160

Chapter 17 Transformations of Categorical Variables 161

17.1 Introduction 161
17.2 General Considerations for Categorical Variables 162
17.3 Derived Variables 164
17.4 Combining Categories 166
17.5 Dummy Coding of Categorical Variables 169
17.6 Multidimensional Categorical Variables 172
17.7 Lookup Tables and External Data 176

Chapter 18 Multiple Interval-Scaled Observations per Subject 179

18.1 Introduction 179
18.2 Static Aggregation 180
18.3 Correlation of Values 184
18.4 Concentration of Values 187
18.5 Course over Time: Standardization of Values 189
18.6 Course over Time: Derived Variables 194

Chapter 19 Multiple Categorical Observations per Subject 201

19.1 Introduction 201
19.2 Absolute and Relative Frequencies of Categories 202
19.3 Concatenating Absolute and Relative Frequencies 206
19.4 Calculating Total and Distinct Counts of the Categories 208
19.5 Using ODS to Create Different Percent Variables 209
19.6 Business Interpretation of Percentage Variables 212
19.7 Other Methods 213

Chapter 20 Coding for Predictive Modeling 215

20.1 Introduction 215
20.2 Proportions or Means of the Target Variable 216
20.3 Interval Variables and Predictive Modeling 225
20.4 Validation Methods 231
20.5 Conclusion 232

Chapter 21 Data Preparation for Multiple-Rows-per-Subject and Longitudinal Data Marts 233

21.1 Introduction 233
21.2 Data Preparation for Association and Sequence Analysis 234

21.3 Enhancing Time Series Data 238

21.4 Aggregating at Various Hierarchical Levels 242

21.5 Preparing Time Series Data with SAS Functions 245

21.6 Using SAS/ETS Procedures for Data Preparation 250

Part 4 Sampling, Scoring, and Automation

Chapter 22 Sampling 257

22.1 Introduction 257

22.2 Sampling Methods 258

22.3 Simple Sampling and Reaching the Exact Sample Count or Proportion 259

22.4 Oversampling 261

22.5 Clustered Sampling 263

22.6 Conclusion 267

Chapter 23 Scoring and Automation 269

23.1 Introduction 269

23.2 Scoring Process 271

23.3 Explicitly Calculating the Score Values from Parameters and Input Variables 273

23.4 Using the Respective SAS/STAT Procedure for Scoring 275

23.5 Scoring with PROC SCORE of SAS/STAT 279

23.6 Using the Respective SAS/ETS Procedure for Scoring 279

23.7 The Score Code That Can Be Produced in SAS Enterprise Miner 281

23.8 The Pre-checks on the Data That Are Useful before Scoring 281

23.9 Automation of Data Mart Creation in General 290

Chapter 24 Do’s and Don’ts When Building Data Marts 293

24.1 Introduction 293

24.2 Process Do’s and Don’ts 294

24.3 Data Mart Handling Do’s and Don’ts 295

24.4 Coding Do’s and Don’ts 298

Part 5 Case Studies

Chapter 25 Case Study 1—Building a Customer Data Mart 305

25.1 Introduction 305

25.2 The Business Context 306

25.3 The Data 307

25.4 The Programs 310

25.5 The Results and Their Usage 316

Chapter 26 Case Study 2—Deriving Customer Segmentation Measures from Transactional Data 317

26.1 The Business Questions 317

26.2 The Data 318

26.3 The Programs 321

26.4 The Results and Their Usage 326

Chapter 27 Case Study 3—Preparing Data for Time Series Analysis 327

27.1 Introduction 327

27.2 The Business Context 328

27.3 The Data 329

27.4 From Transactional Data to the Most Appropriate Aggregation 332

27.5 Comparing PROC SQL, PROC MEANS, and PROC TIMESERIES 336

27.6 Additional Aggregations 338

27.7 Derived Variables 341

27.8 Creating Observations for Future Months 344

27.9 The Results and Their Usage 345

Chapter 28 Case Study 4—Data Preparation in SAS Enterprise Miner 347

28.1 Introduction 347

28.2 Nodes for Data Preparation 348

28.3 Data Definition, Sampling, and Data Partition 350

28.4 Data Transformation 352

28.5 Scoring in SAS Enterprise Miner 352

28.6 Merging Multiple-Rows-per-Subject Data 355

28.7 Conclusion 358

Appendix A Data Structures from a SAS Procedure Point of View 361

A.1 Introduction 361

A.2 Relationship between Data Mart Elements and SAS Procedure Statements 362

A.3 Data Mart Structure Requirements for Selected Base SAS Procedures 363

A.4 Data Mart Structure Requirements for Selected SAS/STAT Procedures 364

A.5 Data Mart Structure Requirements for Selected SAS/ETS Procedures 365

A.6 Data Mart Structure Requirements for Selected SAS/QC Procedures 366

A.7 Data Mart Structure Requirements for Selected SAS/GRAPH Procedures 366

A.8 Data Mart Structure Requirements for SAS Enterprise Miner Nodes 367

Appendix B The Power of SAS for Analytic Data Preparation 369

B.1 Motivation 369

B.2 Overview 370

B.3 Extracting Data from Source Systems 371

B.4 Changing the Data Mart Structure: Transposing 371

B.5 Data Management for Longitudinal and Multiple-Rows-per-Subject Data Sets 372

B.6 Selected Features of the SAS Language for Data Management 375

B.7 Benefits of the SAS Macro Language 378

B.8 Matrix Operations with SAS/IML 381

Appendix C Transposing with DATA Steps 383

C.1 Transposing and Performance 383

C.2 From a LONG Data Set to a WIDE Data Set 385

C.3 From a WIDE Data Set to a LONG Data Set 387

C.4 Transposing Transactional Categories with a DATA Step 389

Glossary 391

Index 393