Contents

Preface *page* xi

PART I INTRODUCTION

1 Motivation 3
1.1 Data Processing and the Research Cycle 4
1.2 What We Do (and Don't Do) in this Book 5
1.3 Why Focus on Data Processing? 7
1.4 Data in Files vs. Data in Databases 9
1.5 Target Audience, Requirements and Software 11
1.6 Plan of the Book 12

2 Gearing Up 14
2.1 R and RStudio 14
2.2 Setting Up the Project Environment for Your Work 16
2.3 The PostgreSQL Database System 20
2.4 Summary and Outlook 22

3 Data = Content + Structure 23
3.1 What Is Data? 23
3.2 Data Content and Structure 24
3.3 Tables, Tables, Tables 26
3.4 The Structure of Tables Matters 30
3.5 Summary and Outlook 35

PART II DATA IN FILES

4 Storing Data in Files 39
4.1 Text and Binary Files 40
4.2 File Formats for Tabular Data 43

Published online by Cambridge University Press

4.3 Transparent and Efficient Use of Files 54
4.4 Summary and Outlook 57

5 Managing Data in Spreadsheets 59
5.1 Application: Spatial Inequality 60
5.2 Spreadsheet Tables and (the Lack of) Structure 63
5.3 Retrieving Data from a Table 64
5.4 Changing Table Structure and Content 66
5.5 Aggregating Data from a Table 67
5.6 Exporting Spreadsheet Data 70
5.7 Results: Spatial Inequality 70
5.8 Summary and Outlook 71

6 Basic Data Management in R 74
6.1 Application: Inequality and Economic Performance in the US 75
6.2 Loading the Data 76
6.3 Merging Tables 79
6.4 Aggregating Data from a Table 82
6.5 Results: Inequality and Economic Performance in the US 84
6.6 Summary and Outlook 85

7 R and the `tidyverse` 87
7.1 Application: Global Patterns of Inequality across Regime Types 88
7.2 A New Operator: The Pipe 89
7.3 Loading the Data 90
7.4 Merging the WID and Polity IV Datasets 92
7.5 Grouping and Aggregation 93
7.6 Results: Global Patterns of Inequality across Regime Types 96
7.7 Other Useful Functions in the `tidyverse` 97
7.8 Summary and Outlook 99

PART III DATA IN DATABASES

8 Introduction to Relational Databases 103
8.1 Database Servers and Clients 105
8.2 SQL Basics 108
8.3 Application: Electoral Disproportionality by Country 109
8.4 Creating a Table with National Elections 110
8.5 Computing Electoral Disproportionality 115
8.6 Results: Electoral Disproportionality by Country 117
8.7 Summary and Outlook 118

9 Relational Databases and Multiple Tables 121
9.1 Application: The Rise of Populism in Europe 122
9.2 Adding the Tables 123
9.3 Joining the Tables 125

Published online by Cambridge University Press

9.4 Merging Data from the PopuList 127
9.5 Maintaining Referential Integrity 129
9.6 Results: The Rise of Populism in Europe 131
9.7 Summary and Outlook 132
10 Database Fine-Tuning 135
10.1 Speeding Up Data Access with Indexes 136
10.2 Collaborative Data Management with Multiple Users 140
10.3 Summary and Outlook 143

PART IV SPECIAL TYPES OF DATA

11 Spatial Data 147
11.1 What Is Spatial Data? 147
11.2 Application: Patterns of Violence in the Bosnian Civil War 150
11.3 Reading and Visualizing Spatial Data in R 151
11.4 Spatial Data in a Relational Database 158
11.5 Results: Patterns of Violence in the Bosnian Civil War 163
11.6 Summary and Outlook 164
12 Text Data 166
12.1 What Is Textual Data? 167
12.2 Application: References to (In)equality in UN Speeches 169
12.3 Working with Strings in (Base) R 170
12.4 Natural Language Processing with `quanteda` 175
12.5 Using PostgreSQL to Manage Documents 179
12.6 Results: References to (In)equality in UN Speeches 183
12.7 Summary and Outlook 184
13 Network Data 187
13.1 What Is Network Data? 187
13.2 Application: Trade and Democracy 190
13.3 Exploring Network Data in R with `igraph` 191
13.4 Network Data in a Relational Database 197
13.5 Results: Trade and Democracy 204
13.6 Summary and Outlook 205

PART V CONCLUSION

14 Best Practices in Data Management 209
14.1 Two General Recommendations 209
14.2 Collaborative Data Management 212
14.3 Disseminating Research Data and Code 214
14.4 Summary and Outlook 216
Bibliography 219
Index 223

Published online by Cambridge University Press

Published online by Cambridge University Press