Contents

Preface xi

Author xiii

1 Data, Exploratory Analysis, and R 1
- 1.1 Why do we analyze data? . . . 1
- 1.2 The view from 90,000 feet . . . 2
  - 1.2.1 Data . . . 2
  - 1.2.2 Exploratory analysis . . . 4
  - 1.2.3 Computers, software, and R . . . 7
- 1.3 A representative R session . . . 11
- 1.4 Organization of this book . . . 21
- 1.5 Exercises . . . 26

2 Graphics in R 29
- 2.1 Exploratory vs. explanatory graphics . . . 29
- 2.2 Graphics systems in R . . . 32
  - 2.2.1 Base graphics . . . 33
  - 2.2.2 Grid graphics . . . 33
  - 2.2.3 Lattice graphics . . . 34
  - 2.2.4 The ggplot2 package . . . 36
- 2.3 The plot function . . . 37
  - 2.3.1 The flexibility of the plot function . . . 37
  - 2.3.2 S3 classes and generic functions . . . 40
  - 2.3.3 Optional parameters for base graphics . . . 42
- 2.4 Adding details to plots . . . 44
  - 2.4.1 Adding points and lines to a scatterplot . . . 44
  - 2.4.2 Adding text to a plot . . . 48
  - 2.4.3 Adding a legend to a plot . . . 49
  - 2.4.4 Customizing axes . . . 50
- 2.5 A few different plot types . . . 52
  - 2.5.1 Pie charts and why they should be avoided . . . 53
  - 2.5.2 Barplot summaries . . . 54
  - 2.5.3 The symbols function . . . 55

2.6 Multiple plot arrays . . . 57
2.6.1 Setting up simple arrays with mfrow . . . 58
2.6.2 Using the layout function . . . 61
2.7 Color graphics . . . 64
2.7.1 A few general guidelines . . . 64
2.7.2 Color options in R . . . 66
2.7.3 The tableplot function . . . 68
2.8 Exercises . . . 70

3 Exploratory Data Analysis: A First Look **79**
3.1 Exploring a new dataset . . . 80
3.1.1 A general strategy . . . 81
3.1.2 Examining the basic data characteristics . . . 82
3.1.3 Variable types in practice . . . 84
3.2 Summarizing numerical data . . . 87
3.2.1 "Typical" values: the mean . . . 88
3.2.2 "Spread": the standard deviation . . . 88
3.2.3 Limitations of simple summary statistics . . . 90
3.2.4 The Gaussian assumption . . . 92
3.2.5 Is the Gaussian assumption reasonable? . . . 95
3.3 Anomalies in numerical data . . . 100
3.3.1 Outliers and their influence . . . 100
3.3.2 Detecting univariate outliers . . . 104
3.3.3 Inliers and their detection . . . 116
3.3.4 Metadata errors . . . 118
3.3.5 Missing data, possibly disguised . . . 120
3.3.6 QQ-plots revisited . . . 125
3.4 Visualizing relations between variables . . . 130
3.4.1 Scatterplots between numerical variables . . . 131
3.4.2 Boxplots: numerical vs. categorical variables . . . 133
3.4.3 Mosaic plots: categorical scatterplots . . . 135
3.5 Exercises . . . 137

4 Working with External Data **141**
4.1 File management in R . . . 142
4.2 Manual data entry . . . 145
4.2.1 Entering the data by hand . . . 145
4.2.2 Manual data entry is bad but sometimes expedient . . . 147
4.3 Interacting with the Internet . . . 148
4.3.1 Previews of three Internet data examples . . . 148
4.3.2 A very brief introduction to HTML . . . 151
4.4 Working with CSV files . . . 152
4.4.1 Reading and writing CSV files . . . 152
4.4.2 Spreadsheets and csv files are *not* the same thing . . . 154
4.4.3 Two potential problems with CSV files . . . 155
4.5 Working with other file types . . . 158

4.5.1 Working with text files . . . 158
4.5.2 Saving and retrieving R objects . . . 162
4.5.3 Graphics files . . . 163
4.6 Merging data from different sources . . . 165
4.7 A brief introduction to databases . . . 168
4.7.1 Relational databases, queries, and SQL . . . 169
4.7.2 An introduction to the `sqldf` package . . . 171
4.7.3 An overview of R's database support . . . 174
4.7.4 An introduction to the `RSQLite` package . . . 175
4.8 Exercises . . . 178

5 Linear Regression Models **181**
5.1 Modeling the `whiteside` data . . . 181
5.1.1 Describing lines in the plane . . . 182
5.1.2 Fitting lines to points in the plane . . . 185
5.1.3 Fitting the `whiteside` data . . . 186
5.2 Overfitting and data splitting . . . 188
5.2.1 An overfitting example . . . 188
5.2.2 The training/validation/holdout split . . . 192
5.2.3 Two useful model validation tools . . . 196
5.3 Regression with multiple predictors . . . 201
5.3.1 The `Cars93` example . . . 202
5.3.2 The problem of collinearity . . . 207
5.4 Using categorical predictors . . . 211
5.5 Interactions in linear regression models . . . 214
5.6 Variable transformations in linear regression . . . 217
5.7 Robust regression: a very brief introduction . . . 221
5.8 Exercises . . . 224

6 Crafting Data Stories **229**
6.1 Crafting good data stories . . . 229
6.1.1 The importance of clarity . . . 230
6.1.2 The basic elements of an effective data story . . . 231
6.2 Different audiences have different needs . . . 232
6.2.1 The executive summary or abstract . . . 233
6.2.2 Extended summaries . . . 234
6.2.3 Longer documents . . . 235
6.3 Three example data stories . . . 235
6.3.1 The Big Mac and Grande Latte economic indices . . . 236
6.3.2 Small losses in the Australian vehicle insurance data . . . 240
6.3.3 Unexpected heterogeneity: the Boston housing data . . . 243

7 Programming in R **247**
7.1 Interactive use versus programming . . . 247
7.1.1 A simple example: computing Fibonnacci numbers . . . 248
7.1.2 Creating your own functions . . . 252
7.2 Key elements of the R language . . . 256
7.2.1 Functions and their arguments . . . 256
7.2.2 The `list` data type . . . 260
7.2.3 Control structures . . . 262
7.2.4 Replacing loops with `apply` functions . . . 268
7.2.5 Generic functions revisited . . . 270
7.3 Good programming practices . . . 275
7.3.1 Modularity and the DRY principle . . . 275
7.3.2 Comments . . . 275
7.3.3 Style guidelines . . . 276
7.3.4 Testing and debugging . . . 276
7.4 Five programming examples . . . 277
7.4.1 The function `ValidationRsquared` . . . 277
7.4.2 The function `TVHsplit` . . . 278
7.4.3 The function `PredictedVsObservedPlot` . . . 278
7.4.4 The function `BasicSummary` . . . 279
7.4.5 The function `FindOutliers` . . . 281
7.5 R scripts . . . 284
7.6 Exercises . . . 285

8 Working with Text Data **289**
8.1 The fundamentals of text data analysis . . . 290
8.1.1 The basic steps in analyzing text data . . . 290
8.1.2 An illustrative example . . . 293
8.2 Basic character functions in R . . . 298
8.2.1 The `nchar` function . . . 298
8.2.2 The `grep` function . . . 301
8.2.3 Application to missing data and alternative spellings . . . 302
8.2.4 The `sub` and `gsub` functions . . . 304
8.2.5 The `strsplit` function . . . 306
8.2.6 Another application: `ConvertAutoMpgRecords` . . . 307
8.2.7 The `paste` function . . . 309
8.3 A brief introduction to regular expressions . . . 311
8.3.1 Regular expression basics . . . 311
8.3.2 Some useful regular expression examples . . . 313
8.4 An aside: ASCII vs. UNICODE . . . 319
8.5 Quantitative text analysis . . . 320
8.5.1 Document-term and document-feature matrices . . . 320
8.5.2 String distances and approximate matching . . . 322
8.6 Three detailed examples . . . 330
8.6.1 Characterizing a book . . . 331
8.6.2 The `cpus` data frame . . . 336

8.6.3 The unclaimed bank account data 344
8.7 Exercises 353

9 Exploratory Data Analysis: A Second Look **357**
9.1 An example: repeated measurements 358
9.1.1 Summary and practical implications 358
9.1.2 The gory details 359
9.2 Confidence intervals and significance 364
9.2.1 Probability models versus data 364
9.2.2 Quantiles of a distribution 366
9.2.3 Confidence intervals 368
9.2.4 Statistical significance and p-values 372
9.3 Characterizing a binary variable 375
9.3.1 The binomial distribution 375
9.3.2 Binomial confidence intervals 377
9.3.3 Odds ratios 382
9.4 Characterizing count data 386
9.4.1 The Poisson distribution and rare events 387
9.4.2 Alternative count distributions 389
9.4.3 Discrete distribution plots 390
9.5 Continuous distributions 393
9.5.1 Limitations of the Gaussian distribution 394
9.5.2 Some alternatives to the Gaussian distribution 398
9.5.3 The `qqPlot` function revisited 404
9.5.4 The problems of ties and implosion 406
9.6 Associations between numerical variables 409
9.6.1 Product-moment correlations 409
9.6.2 Spearman's rank correlation measure 413
9.6.3 The correlation trick 415
9.6.4 Correlation matrices and correlation plots 418
9.6.5 Robust correlations 421
9.6.6 Multivariate outliers 423
9.7 Associations between categorical variables 427
9.7.1 Contingency tables 427
9.7.2 The chi-squared measure and Cramér's V 429
9.7.3 Goodman and Kruskal's tau measure 433
9.8 Principal component analysis (PCA) 438
9.9 Working with date variables 447
9.10 Exercises 449

10 More General Predictive Models **459**
10.1 A predictive modeling overview 459
10.1.1 The predictive modeling problem 460
10.1.2 The model-building process 461
10.2 Binary classification and logistic regression 462
10.2.1 Basic logistic regression formulation 462

10.2.2 Fitting logistic regression models . . . 464
10.2.3 Evaluating binary classifier performance . . . 467
10.2.4 A brief introduction to glms . . . 474
10.3 Decision tree models . . . 478
10.3.1 Structure and fitting of decision trees . . . 479
10.3.2 A classification tree example . . . 485
10.3.3 A regression tree example . . . 487
10.4 Combining trees with regression . . . 491
10.5 Introduction to machine learning models . . . 498
10.5.1 The instability of simple tree-based models . . . 499
10.5.2 Random forest models . . . 500
10.5.3 Boosted tree models . . . 502
10.6 Three practical details . . . 506
10.6.1 Partial dependence plots . . . 507
10.6.2 Variable importance measures . . . 513
10.6.3 Thin levels and data partitioning . . . 519
10.7 Exercises . . . 521

11 Keeping It All Together **525**
11.1 Managing your *R* installation . . . 525
11.1.1 Installing *R* . . . 526
11.1.2 Updating packages . . . 526
11.1.3 Updating *R* . . . 527
11.2 Managing files effectively . . . 528
11.2.1 Organizing directories . . . 528
11.2.2 Use appropriate file extensions . . . 531
11.2.3 Choose good file names . . . 532
11.3 Document everything . . . 533
11.3.1 Data dictionaries . . . 533
11.3.2 Documenting code . . . 534
11.3.3 Documenting results . . . 535
11.4 Introduction to reproducible computing . . . 536
11.4.1 The key ideas of reproducibility . . . 536
11.4.2 Using R Markdown . . . 537

Bibliography **539**

Index **544**