Contents

Preface xxv

Acknowledgments xxvii

Introduction xxix

1 Getting started with R and RStudio 1

1.1 Why R? 1

1.2 The R console 1

1.3 Scripts 2

1.4 RStudio 3

1.4.1 The panes 3

1.4.2 Key bindings 5

1.4.3 Running commands while editing scripts 6

1.4.4 Changing global options 8

1.5 Installing R packages 8

I R 11

2 R basics 13

2.1 Case study: US Gun Murders 13

2.2 The very basics 15

2.2.1 Objects 15

2.2.2 The workspace 15

2.2.3 Functions 16

2.2.4 Other prebuilt objects 18

2.2.5 Variable names 18

2.2.6 Saving your workspace 19

2.2.7 Motivating scripts 19

2.2.8 Commenting your code 20

2.3 Exercises 20

2.4 Data types 21

2.4.1 Data frames 21

2.4.2 Examining an object 21

2.4.3 The accessor: $ 22

2.4.4 Vectors: numerics, characters, and logical 23

2.4.5 Factors 24

2.4.6 Lists 24

2.4.7 Matrices 26

2.5 Exercises 27

2.6 Vectors 28

2.6.1 Creating vectors 28

2.6.2 Names 29

2.6.3 Sequences 29

2.6.4 Subsetting 30

2.7 Coercion 31

2.7.1 Not availables (NA) 32

2.8 Exercises 32

2.9 Sorting 33

2.9.1 `sort` 33

2.9.2 `order` 33

2.9.3 `max` and `which.max` 34

2.9.4 `rank` 34

2.9.5 Beware of recycling 35

2.10 Exercises 35

2.11 Vector arithmetics 36

2.11.1 Rescaling a vector 37

2.11.2 Two vectors 37

2.12 Exercises 38

2.13 Indexing 38

2.13.1 Subsetting with logicals 38

2.13.2 Logical operators 39

2.13.3 `which` 39

2.13.4 `match` 40

2.13.5 `%in%` 40

2.14 Exercises 40
2.15 Basic plots 41
2.15.1 plot 41
2.15.2 hist 42
2.15.3 boxplot 43
2.15.4 image 43
2.16 Exercises 44

3 Programming basics 45
3.1 Conditional expressions 45
3.2 Defining functions 47
3.3 Namespaces 48
3.4 For-loops 49
3.5 Vectorization and functionals 50
3.6 Exercises 51

4 The tidyverse 53
4.1 Tidy data 53
4.2 Exercises 54
4.3 Manipulating data frames 55
4.3.1 Adding a column with `mutate` 55
4.3.2 Subsetting with `filter` 56
4.3.3 Selecting columns with `select` 56
4.4 Exercises 57
4.5 The pipe: %>% 58
4.6 Exercises 59
4.7 Summarizing data 60
4.7.1 summarize 60
4.7.2 pull 62
4.7.3 Group then summarize with `group_by` 63
4.8 Sorting data frames 64
4.8.1 Nested sorting 64
4.8.2 The top n 65
4.9 Exercises 65
4.10 Tibbles 66

4.10.1 Tibbles display better 67
4.10.2 Subsets of tibbles are tibbles 67
4.10.3 Tibbles can have complex entries 68
4.10.4 Tibbles can be grouped 68
4.10.5 Create a tibble using `tibble` instead of `data.frame` 68
4.11 The dot operator 69
4.12 `do` 70
4.13 The **purrr** package 71
4.14 Tidyverse conditionals 73
4.14.1 `case_when` 73
4.14.2 `between` 73
4.15 Exercises 74
5 Importing data **75**
5.1 Paths and the working directory 76
5.1.1 The filesystem 76
5.1.2 Relative and full paths 77
5.1.3 The working directory 77
5.1.4 Generating path names 78
5.1.5 Copying files using paths 78
5.2 The readr and readxl packages 79
5.2.1 readr 79
5.2.2 readxl 80
5.3 Exercises 80
5.4 Downloading files 81
5.5 R-base importing functions 82
5.5.1 `scan` 82
5.6 Text versus binary files 83
5.7 Unicode versus ASCII 83
5.8 Organizing data with spreadsheets 84
5.9 Exercises 84

II Data Visualization 85

6 Introduction to data visualization 87

7 ggplot2 91
7.1 The components of a graph . . . 92
7.2 ggplot objects . . . 93
7.3 Geometries . . . 94
7.4 Aesthetic mappings . . . 95
7.5 Layers . . . 96
7.5.1 Tinkering with arguments . . . 97
7.6 Global versus local aesthetic mappings . . . 98
7.7 Scales . . . 99
7.8 Labels and titles . . . 100
7.9 Categories as colors . . . 101
7.10 Annotation, shapes, and adjustments . . . 102
7.11 Add-on packages . . . 103
7.12 Putting it all together . . . 104
7.13 Quick plots with qplot . . . 105
7.14 Grids of plots . . . 106
7.15 Exercises . . . 106

8 Visualizing data distributions 109
8.1 Variable types . . . 109
8.2 Case study: describing student heights . . . 110
8.3 Distribution function . . . 110
8.4 Cumulative distribution functions . . . 111
8.5 Histograms . . . 112
8.6 Smoothed density . . . 113
8.6.1 Interpreting the y-axis . . . 117
8.6.2 Densities permit stratification . . . 118
8.7 Exercises . . . 118
8.8 The normal distribution . . . 122
8.9 Standard units . . . 124
8.10 Quantile-quantile plots . . . 125
8.11 Percentiles . . . 127

8.12 Boxplots 127
8.13 Stratification 129
8.14 Case study: describing student heights (continued) 129
8.15 Exercises 131
8.16 ggplot2 geometries 132
8.16.1 Barplots 133
8.16.2 Histograms 134
8.16.3 Density plots 135
8.16.4 Boxplots 136
8.16.5 QQ-plots 136
8.16.6 Images 137
8.16.7 Quick plots 138
8.17 Exercises 140
9 Data visualization in practice **141**
9.1 Case study: new insights on poverty 141
9.1.1 Hans Rosling's quiz 142
9.2 Scatterplots 143
9.3 Faceting 144
9.3.1 `facet_wrap` 146
9.3.2 Fixed scales for better comparisons 147
9.4 Time series plots 147
9.4.1 Labels instead of legends 150
9.5 Data transformations 151
9.5.1 Log transformation 151
9.5.2 Which base? 153
9.5.3 Transform the values or the scale? 154
9.6 Visualizing multimodal distributions 155
9.7 Comparing multiple distributions with boxplots and ridge plots 155
9.7.1 Boxplots 156
9.7.2 Ridge plots 157
9.7.3 Example: 1970 versus 2010 income distributions 159
9.7.4 Accessing computed variables 164
9.7.5 Weighted densities 167
9.8 The ecological fallacy and importance of showing the data 167

9.8.1 Logistic transformation 168
9.8.2 Show the data 168

10 Data visualization principles 171
10.1 Encoding data using visual cues 171
10.2 Know when to include 0 174
10.3 Do not distort quantities 177
10.4 Order categories by a meaningful value 179
10.5 Show the data 180
10.6 Ease comparisons 183
10.6.1 Use common axes 183
10.6.2 Align plots vertically to see horizontal changes and horizontally to see vertical changes 184
10.6.3 Consider transformations 185
10.6.4 Visual cues to be compared should be adjacent 187
10.6.5 Use color 188
10.7 Think of the color blind 188
10.8 Plots for two variables 189
10.8.1 Slope charts 189
10.8.2 Bland-Altman plot 191
10.9 Encoding a third variable 191
10.10 Avoid pseudo-three-dimensional plots 193
10.11 Avoid too many significant digits 195
10.12 Know your audience 196
10.13 Exercises 196
10.14 Case study: vaccines and infectious diseases 201
10.15 Exercises 204

11 Robust summaries 205
11.1 Outliers 205
11.2 Median 206
11.3 The inter quartile range (IQR) 206
11.4 Tukey's definition of an outlier 207
11.5 Median absolute deviation 208
11.6 Exercises 208
11.7 Case study: self-reported student heights 209

III Statistics with R 213

12 Introduction to statistics with R 215

13 Probability 217

13.1 Discrete probability 217

13.1.1 Relative frequency 217

13.1.2 Notation 218

13.1.3 Probability distributions 218

13.2 Monte Carlo simulations for categorical data 218

13.2.1 Setting the random seed 220

13.2.2 With and without replacement 220

13.3 Independence 221

13.4 Conditional probabilities 221

13.5 Addition and multiplication rules 222

13.5.1 Multiplication rule 222

13.5.2 Multiplication rule under independence 222

13.5.3 Addition rule 223

13.6 Combinations and permutations 223

13.6.1 Monte Carlo example 227

13.7 Examples 227

13.7.1 Monty Hall problem 228

13.7.2 Birthday problem 229

13.8 Infinity in practice 231

13.9 Exercises 232

13.10 Continuous probability 234

13.11 Theoretical continuous distributions 235

13.11.1 Theoretical distributions as approximations 235

13.11.2 The probability density 237

13.12 Monte Carlo simulations for continuous variables 238

13.13 Continuous distributions 239

13.14 Exercises 239

14 Random variables 241

14.1 Random variables 241

14.2 Sampling models 242

14.3 The probability distribution of a random variable 243
14.4 Distributions versus probability distributions 245
14.5 Notation for random variables 245
14.6 The expected value and standard error 246
14.6.1 Population SD versus the sample SD 248
14.7 Central Limit Theorem 249
14.7.1 How large is large in the Central Limit Theorem? 250
14.8 Statistical properties of averages 250
14.9 Law of large numbers 252
14.9.1 Misinterpreting law of averages 252
14.10 Exercises 252
14.11 Case study: The Big Short 254
14.11.1 Interest rates explained with chance model 254
14.11.2 The Big Short 257
14.12 Exercises 260

15 Statistical inference **261**

15.1 Polls 261
15.1.1 The sampling model for polls 262
15.2 Populations, samples, parameters, and estimates 264
15.2.1 The sample average 264
15.2.2 Parameters 265
15.2.3 Polling versus forecasting 265
15.2.4 Properties of our estimate: expected value and standard error . . 266
15.3 Exercises 267
15.4 Central Limit Theorem in practice 268
15.4.1 A Monte Carlo simulation 269
15.4.2 The spread 271
15.4.3 Bias: why not run a very large poll? 271
15.5 Exercises 272
15.6 Confidence intervals 274
15.6.1 A Monte Carlo simulation 276
15.6.2 The correct language 277
15.7 Exercises 277
15.8 Power 278

15.9 p-values . . . 279
15.10 Association tests . . . 280
15.10.1 Lady Tasting Tea . . . 281
15.10.2 Two-by-two tables . . . 282
15.10.3 Chi-square Test . . . 282
15.10.4 The odds ratio . . . 283
15.10.5 Confidence intervals for the odds ratio . . . 284
15.10.6 Small count correction . . . 285
15.10.7 Large samples, small p-values . . . 285
15.11 Exercises . . . 286

16 Statistical models **287**
16.1 Poll aggregators . . . 288
16.1.1 Poll data . . . 290
16.1.2 Pollster bias . . . 292
16.2 Data-driven models . . . 293
16.3 Exercises . . . 295
16.4 Bayesian statistics . . . 298
16.4.1 Bayes theorem . . . 298
16.5 Bayes theorem simulation . . . 299
16.5.1 Bayes in practice . . . 300
16.6 Hierarchical models . . . 301
16.7 Exercises . . . 303
16.8 Case study: election forecasting . . . 305
16.8.1 Bayesian approach . . . 306
16.8.2 The general bias . . . 307
16.8.3 Mathematical representations of models . . . 307
16.8.4 Predicting the electoral college . . . 310
16.8.5 Forecasting . . . 314
16.9 Exercises . . . 317
16.10 The t-distribution . . . 318

17 Regression **321**
17.1 Case study: is height hereditary? . . . 321
17.2 The correlation coefficient . . . 322

17.2.1 Sample correlation is a random variable 324
17.2.2 Correlation is not always a useful summary 326
17.3 Conditional expectations 326
17.4 The regression line 329
17.4.1 Regression improves precision 330
17.4.2 Bivariate normal distribution (advanced) 331
17.4.3 Variance explained 333
17.4.4 Warning: there are two regression lines 333
17.5 Exercises 334

18 Linear models **335**

18.1 Case study: Moneyball 335
18.1.1 Sabermetics 336
18.1.2 Baseball basics 337
18.1.3 No awards for BB 338
18.1.4 Base on balls or stolen bases? 339
18.1.5 Regression applied to baseball statistics 341
18.2 Confounding 344
18.2.1 Understanding confounding through stratification 345
18.2.2 Multivariate regression 348
18.3 Least squares estimates 348
18.3.1 Interpreting linear models 349
18.3.2 Least Squares Estimates (LSE) 349
18.3.3 The `lm` function 351
18.3.4 LSE are random variables 352
18.3.5 Predicted values are random variables 353
18.4 Exercises 354
18.5 Linear regression in the tidyverse 355
18.5.1 The broom package 358
18.6 Exercises 359
18.7 Case study: Moneyball (continued) 360
18.7.1 Adding salary and position information 364
18.7.2 Picking nine players 365
18.8 The regression fallacy 367
18.9 Measurement error models 369

18.10 Exercises . . . 371

19 Association is not causation 373

19.1 Spurious correlation . . . 373
19.2 Outliers . . . 376
19.3 Reversing cause and effect . . . 378
19.4 Confounders . . . 379
19.4.1 Example: UC Berkeley admissions . . . 379
19.4.2 Confounding explained graphically . . . 380
19.4.3 Average after stratifying . . . 381
19.5 Simpson's paradox . . . 382
19.6 Exercises . . . 383

IV Data Wrangling 385

20 Introduction to data wrangling 387

21 Reshaping data 389

21.1 gather . . . 389
21.2 spread . . . 391
21.3 separate . . . 391
21.4 unite . . . 394
21.5 Exercises . . . 395

22 Joining tables 397

22.1 Joins . . . 398
22.1.1 Left join . . . 399
22.1.2 Right join . . . 400
22.1.3 Inner join . . . 400
22.1.4 Full join . . . 400
22.1.5 Semi join . . . 401
22.1.6 Anti join . . . 401
22.2 Binding . . . 402
22.2.1 Binding columns . . . 402
22.2.2 Binding by rows . . . 402
22.3 Set operators . . . 403
22.3.1 Intersect . . . 403

22.3.2 Union 404

22.3.3 `setdiff` 404

22.3.4 `setequal` 404

22.4 Exercises 405

23 Web scraping **407**

23.1 HTML 408

23.2 The rvest package 409

23.3 CSS selectors 411

23.4 JSON 412

23.5 Exercises 413

24 String processing **415**

24.1 The stringr package 415

24.2 Case study 1: US murders data 417

24.3 Case study 2: self-reported heights 419

24.4 How to *escape* when defining strings 421

24.5 Regular expressions 423

24.5.1 Strings are a regexp 423

24.5.2 Special characters 423

24.5.3 Character classes 425

24.5.4 Anchors 426

24.5.5 Quantifiers 426

24.5.6 White space \s 427

24.5.7 Quantifiers: *, ?, + ... 428

24.5.8 Not 428

24.5.9 Groups 429

24.6 Search and replace with regex 430

24.6.1 Search and replace using groups 432

24.7 Testing and improving 433

24.8 Trimming 435

24.9 Changing lettercase 436

24.10 Case study 2: self-reported heights (continued) 436

24.10.1 The `extract` function 437

24.10.2 Putting it all together 438

24.11 String splitting 439
24.12 Case study 3: extracting tables from a PDF 442
24.13 Recoding 445
24.14 Exercises 446

25 Parsing dates and times **449**
25.1 The date data type 449
25.2 The lubridate package 450
25.3 Exercises 453

26 Text mining **455**
26.1 Case study: Trump tweets 455
26.2 Text as data 457
26.3 Sentiment analysis 462
26.4 Exercises 467

V Machine Learning **469**

27 Introduction to machine learning **471**
27.1 Notation 471
27.2 An example 472
27.3 Exercises 474
27.4 Evaluation metrics 474
27.4.1 Training and test sets 475
27.4.2 Overall accuracy 476
27.4.3 The confusion matrix 478
27.4.4 Sensitivity and specificity 479
27.4.5 Balanced accuracy and F_1 score 481
27.4.6 Prevalence matters in practice 482
27.4.7 ROC and precision-recall curves 483
27.4.8 The loss function 484
27.5 Exercises 486
27.6 Conditional probabilities and expectations 486
27.6.1 Conditional probabilities 487
27.6.2 Conditional expectations 488
27.6.3 Conditional expectation minimizes squared loss function 488
27.7 Exercises 489

27.8 Case study: is it a 2 or a 7? 489

28 Smoothing **493**

28.1 Bin smoothing 495

28.2 Kernels 497

28.3 Local weighted regression (loess) 498

28.3.1 Fitting parabolas 502

28.3.2 Beware of default smoothing parameters 503

28.4 Connecting smoothing to machine learning 504

28.5 Exercises 504

29 Cross validation **507**

29.1 Motivation with k-nearest neighbors 507

29.1.1 Over-training 509

29.1.2 Over-smoothing 510

29.1.3 Picking the k in kNN 511

29.2 Mathematical description of cross validation 513

29.3 K-fold cross validation 514

29.4 Exercises 517

29.5 Bootstrap 518

29.6 Exercises 521

30 The caret package **523**

30.1 The caret `train` functon 523

30.2 Cross validation 524

30.3 Example: fitting with loess 526

31 Examples of algorithms **529**

31.1 Linear regression 529

31.1.1 The `predict` function 530

31.2 Exercises 531

31.3 Logistic regression 533

31.3.1 Generalized linear models 534

31.3.2 Logistic regression with more than one predictor 538

31.4 Exercises 539

31.5 k-nearest neighbors 540

31.6 Exercises 541

31.7 Generative models . . . 541
31.7.1 Naive Bayes . . . 542
31.7.2 Controlling prevalence . . . 543
31.7.3 Quadratic discriminant analysis . . . 545
31.7.4 Linear discriminant analysis . . . 547
31.7.5 Connection to distance . . . 549
31.8 Case study: more than three classes . . . 549
31.9 Exercises . . . 553
31.10 Classification and regression trees (CART) . . . 554
31.10.1 The curse of dimensionality . . . 554
31.10.2 CART motivation . . . 555
31.10.3 Regression trees . . . 558
31.10.4 Classification (decision) trees . . . 564
31.11 Random forests . . . 566
31.12 Exercises . . . 571

32 Machine learning in practice 573

32.1 Preprocessing . . . 574
32.2 k-nearest neighbor and random forest . . . 575
32.3 Variable importance . . . 578
32.4 Visual assessments . . . 579
32.5 Ensembles . . . 579
32.6 Exercises . . . 580

33 Large datasets 581

33.1 Matrix algebra . . . 581
33.1.1 Notation . . . 582
33.1.2 Converting a vector to a matrix . . . 584
33.1.3 Row and column summaries . . . 585
33.1.4 apply . . . 586
33.1.5 Filtering columns based on summaries . . . 586
33.1.6 Indexing with matrices . . . 588
33.1.7 Binarizing the data . . . 590
33.1.8 Vectorization for matrices . . . 590
33.1.9 Matrix algebra operations . . . 591

33.2 Exercises . . . 591
33.3 Distance . . . 591
33.3.1 Euclidean distance . . . 592
33.3.2 Distance in higher dimensions . . . 592
33.3.3 Euclidean distance example . . . 593
33.3.4 Predictor space . . . 595
33.3.5 Distance between predictors . . . 595
33.4 Exercises . . . 595
33.5 Dimension reduction . . . 596
33.5.1 Preserving distance . . . 596
33.5.2 Linear transformations (advanced) . . . 599
33.5.3 Orthogonal transformations (advanced) . . . 600
33.5.4 Principal component analysis . . . 602
33.5.5 Iris example . . . 604
33.5.6 MNIST example . . . 607
33.6 Exercises . . . 609
33.7 Recommendation systems . . . 610
33.7.1 Movielens data . . . 610
33.7.2 Recommendation systems as a machine learning challenge . . . 612
33.7.3 Loss function . . . 612
33.7.4 A first model . . . 613
33.7.5 Modeling movie effects . . . 614
33.7.6 User effects . . . 615
33.8 Exercises . . . 616
33.9 Regularization . . . 617
33.9.1 Motivation . . . 617
33.9.2 Penalized least squares . . . 619
33.9.3 Choosing the penalty terms . . . 622
33.10 Exercises . . . 624
33.11 Matrix factorization . . . 625
33.11.1 Factors analysis . . . 628
33.11.2 Connection to SVD and PCA . . . 630
33.12 Exercises . . . 633

34 Clustering 639
34.1 Hierarchical clustering 640
34.2 k-means 642
34.3 Heatmaps 642
34.4 Filtering features 643
34.5 Exercises 644

VI Productivity Tools 645

35 Introduction to productivity tools 647

36 Organizing with Unix 649
36.1 Naming convention 649
36.2 The terminal 650
36.3 The filesystem 650
36.3.1 Directories and subdirectories 651
36.3.2 The home directory 651
36.3.3 Working directory 652
36.3.4 Paths 653
36.4 Unix commands 653
36.4.1 `ls`: Listing directory content 654
36.4.2 `mkdir` and `rmdir`: make and remove a directory 654
36.4.3 `cd`: navigating the filesystem by changing directories 655
36.5 Some examples 657
36.6 More Unix commands 658
36.6.1 `mv`: moving files 658
36.6.2 `cp`: copying files 659
36.6.3 `rm`: removing files 659
36.6.4 `less`: looking at a file 659
36.7 Preparing for a data science project 660
36.8 Advanced Unix 661
36.8.1 Arguments 661
36.8.2 Getting help 662
36.8.3 Pipes 662
36.8.4 Wild cards 663
36.8.5 Environment variables 663

36.8.6 Shells 664

36.8.7 Executables 664

36.8.8 Permissions and file types 665

36.8.9 Commands you should learn 665

36.8.10 File manipulation in R 665

37 Git and GitHub **667**

37.1 Why use Git and GitHub? 667

37.2 GitHub accounts 667

37.3 GitHub repositories 670

37.4 Overview of Git 671

37.4.1 Clone 672

37.5 Initializing a Git directory 676

37.6 Using Git and GitHub in RStudio 678

38 Reproducible projects with RStudio and R markdown **683**

38.1 RStudio projects 683

38.2 R markdown 686

38.2.1 The header 688

38.2.2 R code chunks 688

38.2.3 Global options 689

38.2.4 knitR 689

38.2.5 More on R markdown 690

38.3 Organizing a data science project 690

38.3.1 Create directories in Unix 690

38.3.2 Create an RStudio project 691

38.3.3 Edit some R scripts 692

38.3.4 Create some more directories using Unix 693

38.3.5 Add a README file 693

38.3.6 Initializing a Git directory 693

38.3.7 Add, commit, and push files using RStudio 694

Index **695**