Contents

Chapter 1: Introduction 1

1.1 What are insights? 1
   1.1.1 Dictionary 1
   1.1.2 The business perspective 2
   1.1.3 Our definition 3
   1.1.4 Our ecology example . . . We love fruit 3
1.2 Question, question, question (how are data born?) 5
1.3 But what exactly are data? 7
1.4 Response and predictor variables 8
1.5 Some key features of datasets 9
1.6 Demonstrations of getting insights from data 11
1.7 The general Insights workflow 16
1.8 Summing up and looking forward 17

Chapter 2: Getting acquainted 19

2.1 Getting acquainted with R and RStudio 19
   2.1.1 Why R? 20
   2.1.2 Why RStudio? 21
   2.1.3 Getting and installing R 22
   2.1.4 Getting and installing RStudio 23
   2.1.5 A brief tour of RStudio 24
2.2 Your first R command! 26
   2.2.1 Getting to know R a little better 27
   2.2.2 Storing and reusing results 29
   2.2.3 What names should I use? 31
2.3 Writing scripts 32
   2.3.1 Comments in your scripts 34
   2.3.2 Save and keep safe your script file 35
   2.3.3 Running your scripts 36
2.4 When things go wrong . . . 36
   2.4.1 Errors 37
2.4.2 Warnings 38
2.4.3 The dreaded + 38
2.5 Functions 39
  2.5.1 Functions, the sequel 41
2.6 Add-on packages 42
  2.6.1 Finding add-on packages 43
  2.6.2 Installing (downloading) packages 44
  2.6.3 Loading packages 46
  2.6.4 An analogy 46
  2.6.5 Updating R, RStudio, and your packages 47
2.7 Getting help 48
  2.7.1 R help system and files 48
  2.7.2 Navigating help files 49
  2.7.3 Vignettes 50
  2.7.4 Cheat sheets 50
  2.7.5 Other sources of help 51
  2.7.6 Asking for help from others 51
2.8 Common pitfalls 52
2.9 Summing up and looking forward 52

Chapter 3: Workflow Demonstration part 1: Preparation 55
3.1 What is the question? 57
  3.1.1 The three response variables 58
  3.1.2 The hypotheses 59
3.2 Design of the study 60
3.3 Preparing your data 61
  3.3.1 Acquire the dataset 64
3.4 Preparing your computer 66
  3.4.1 Making the project folder for the bat data 67
  3.4.2 Projects in RStudio 68
  3.4.3 Create a new R script and load packages 71
3.5 Get the data into R 72
  3.5.1 View and refine the import 76
3.6 Getting going with data management 78
  3.6.1 How the data are stored in R 79
3.7 Clean and tidy the data 81
  3.7.1 Tidying the data 82
  3.7.2 Cleaning the data 82
  3.7.3 Refine the variable names 83
  3.7.4 Fix the dates 85
  3.7.5 Rename some values in a variable 86
  3.7.6 Check for duplicates 87
CONTENTS

3.7.7 Check for implausible and invalid values 89
3.7.8 What about those NAs? 90
3.8 Stop that! Don't even think about it! 92
  3.8.1 Don't mess with the 'working directory' 92
  3.8.2 Don't use the data import tool or file.choose 93
  3.8.3 Don't even think about using the attach function 93
  3.8.4 Avoid using square brackets or dollar signs 93
3.9 Summing up and looking forward 94

Chapter 4: Workflow Demonstration part 2: Getting insights 97
  4.1 Initial insights 1: Numbers and counting 98
    4.1.1 Our first insights: The number, sex, and age of bats 98
  4.2 Initial insights 2: Distributions 103
    4.2.1 Insights . . . . you've done it! 105
  4.3 Transform the data 108
  4.4 Insights about our questions 111
    4.4.1 Distribution of number of prey 111
    4.4.2 Shapes: Mean wingspan 113
    4.4.3 Shapes: Proportion migratory 114
    4.4.4 Relationships 116
    4.4.5 Communication (beautifying the graphs) 121
    4.4.6 Beautifying the wingspan, age, sex graph 122
  4.5 Another view of the question and data 125
    4.5.1 Before you continue . . . 125
    4.5.2 A prey-centric view 125
  4.6 A caveat 138
  4.7 Summing up and looking forward 138
  4.8 A small reward, if you like dogs 139

Chapter 5: Dealing with data 1: Digging into dplyr 141
  5.1 Introducing dplyr 142
    5.1.1 Selecting variables with the select function 143
    5.1.2 Renaming variables with select and rename 146
    5.1.3 Creating new variables with the mutate function 146
    5.1.4 Getting particular observations with filter 149
    5.1.5 Ordering observations with arrange 153
  5.2 Grouping and summarizing data with dplyr 155
    5.2.1 Summarizing data—the nitty-gritty 156
    5.2.2 Grouped summaries using group_by magic 160
    5.2.3 More than one grouping variable 163
    5.2.4 Using group_by with other verbs 165
## 5.2.5 Removing grouping information

5.3 Summing up and looking forward

### Chapter 6: Dealing with data 2: Expanding your toolkit

6.1 Pipes and pipelines
   6.1.1 Why do we need pipes?
   6.1.2 On why you shouldn’t nest functions

6.2 Subduing the pesky string

6.3 Elegantly managing dates and times
   6.3.1 Date/time formats
   6.3.2 Dates in the hat project data
   6.3.3 Why parse dates?
   6.3.4 More about parsing dates/times
   6.3.5 Calculations with dates/times

6.4 Changing between wider and longer data arrangements
   6.4.1 Going longer
   6.4.2 Going wider

6.5 Summing up and looking forward

### Chapter 7: Getting to grips with ggplot2

7.1 Anatomy of a ggplot
   7.1.1 Layers
   7.1.2 Scales
   7.1.3 Coordinate system
   7.1.4 Fantastic faceting

7.2 Putting it into practice
   7.2.1 Inheriting data and aesthetics from ggplot

7.3 Beautifying plots
   7.3.1 Working with layer-specific geom properties
   7.3.2 Adding titles and labels
   7.3.3 Themes

7.4 Summing up and looking forward

### Chapter 8: Making deeper insights part 1: Working with single variables

8.1 Variables and data
   8.1.1 Numeric versus categorical variables
   8.1.2 Ratio versus interval scales

8.2 Samples and distributions
   8.2.1 Understanding numerical variables

8.3 Graphical summaries of numeric variables
   8.3.1 Making some insights about wingspan
CONTENTS

8.3.2 Descriptive statistics for numeric variables 227
8.3.3 Measuring central tendency 228
8.3.4 Measuring dispersion 229
8.3.5 Mapping measures of central tendency and dispersion to a figure 231
8.3.6 Combining histograms and boxplots 233
8.4 A moment with missing values in numeric variables (NAs) 234
8.5 Exploring a categorical variable 236
8.5.1 Understanding categorical variables 236
8.6 Summing up and looking forward 244
8.7 A cat-related reward 245

Chapter 9: Making deeper insights part 1: Relationships among (many) variables 247

9.1 Associations between two numeric variables 248
9.1.1 Descriptive statistics: Correlations 248
9.1.2 Other measures of correlation 251
9.1.3 Graphical summaries between two numeric variables:
   The scatterplot 252
9.2 Associations between two categorical variables 256
9.2.1 Numerical summaries 256
9.2.2 Graphical summaries 258
9.2.3 An alternative, and perhaps more valuable 260
9.3 Categorical–numerical associations 261
9.3.1 Numerical summaries 262
9.3.2 Graphical summaries for numerical versus categorical data 262
9.3.3 Alternatives to box-and-whisker plots 264
9.4 Building in complexity: Relationships among three or more variables 267
9.5 Summing up and looking forward 269

Chapter 10: Looking back and looking forward 271

10.1 Next learning steps 272
10.2 Reproducibility: What, why, and how? 274
   10.2.1 Why should you try and make your work reproducible? 274
   10.2.2 How can you make your work more reproducible? 275
10.3 Congratulations! 281

Index 283