

INTRODUCTION TO DATA SCIENCE

Contents

Preface	23
Acknowledgments	25
Introduction	27
1 Getting started with R and RStudio	29
1.1 Why R?	29
1.2 The R console	29
1.3 Scripts	30
1.4 RStudio	31
1.4.1 The panes	31
1.4.2 Key bindings	33
1.4.3 Running commands while editing scripts	34
1.4.4 Changing global options	36
1.5 Installing R packages	36
I R	39
2 R basics	41
2.1 Case study: US Gun Murders	41
2.2 The very basics	43
2.2.1 Objects	43
2.2.2 The workspace	43
2.2.3 Functions	44
2.2.4 Other prebuilt objects	46
2.2.5 Variable names	47
2.2.6 Saving your workspace	47
2.2.7 Motivating scripts	47
2.2.8 Commenting your code	48

2.3	Exercises	48
2.4	Data types	49
2.4.1	Data frames	49
2.4.2	Examining an object	49
2.4.3	The accessor: <code>\$</code>	50
2.4.4	Vectors: numerics, characters, and logical	51
2.4.5	Factors	52
2.4.6	Lists	52
2.4.7	Matrices	54
2.5	Exercises	55
2.6	Vectors	56
2.6.1	Creating vectors	56
2.6.2	Names	57
2.6.3	Sequences	58
2.6.4	Subsetting	58
2.7	Coercion	59
2.7.1	Not availables (NA)	60
2.8	Exercises	60
2.9	Sorting	61
2.9.1	<code>sort</code>	61
2.9.2	<code>order</code>	61
2.9.3	<code>max</code> and <code>which.max</code>	62
2.9.4	<code>rank</code>	63
2.9.5	Beware of recycling	63
2.10	Exercises	64
2.11	Vector arithmetics	65
2.11.1	Rescaling a vector	65
2.11.2	Two vectors	65
2.12	Exercises	66
2.13	Indexing	66
2.13.1	Subsetting with logicals	67
2.13.2	Logical operators	67
2.13.3	<code>which</code>	68
2.13.4	<code>match</code>	68

<i>0.0 Contents</i>	5
2.13.5 <code>%in%</code>	69
2.14 Exercises	69
2.15 Basic plots	70
2.15.1 <code>plot</code>	70
2.15.2 <code>hist</code>	70
2.15.3 <code>boxplot</code>	71
2.15.4 <code>image</code>	72
2.16 Exercises	72
3 Programming basics	73
3.1 Conditional expressions	73
3.2 Defining functions	75
3.3 Namespaces	76
3.4 For-loops	77
3.5 Vectorization and functionals	78
3.6 Exercises	79
4 The tidyverse	81
4.1 Tidy data	81
4.2 Exercises	82
4.3 Manipulating data frames	83
4.3.1 Adding a column with <code>mutate</code>	83
4.3.2 Subsetting with <code>filter</code>	84
4.3.3 Selecting columns with <code>select</code>	84
4.4 Exercises	85
4.5 The pipe: <code>%>%</code>	86
4.6 Exercises	87
4.7 Summarizing data	88
4.7.1 <code>summarize</code>	88
4.7.2 <code>pull</code>	90
4.7.3 Group then summarize with <code>group_by</code>	91
4.8 Sorting data frames	92
4.8.1 Nested sorting	93
4.8.2 The top <i>n</i>	93
4.9 Exercises	93

4.10	Tibbles	95
4.10.1	Tibbles display better	95
4.10.2	Subsets of tibbles are tibbles	96
4.10.3	Tibbles can have complex entries	96
4.10.4	Tibbles can be grouped	97
4.10.5	Create a tibble using <code>tibble</code> instead of <code>data.frame</code>	97
4.11	The dot operator	97
4.12	<code>do</code>	98
4.13	The <code>purrr</code> package	100
4.14	Tidyverse conditionals	101
4.14.1	<code>case_when</code>	101
4.14.2	<code>between</code>	102
4.15	Exercises	102
5	Importing data	103
5.1	Paths and the working directory	104
5.1.1	The filesystem	104
5.1.2	Relative and full paths	105
5.1.3	The working directory	105
5.1.4	Generating path names	106
5.1.5	Copying files using paths	106
5.2	The <code>readr</code> and <code>readxl</code> packages	107
5.2.1	<code>readr</code>	107
5.2.2	<code>readxl</code>	108
5.3	Exercises	108
5.4	Downloading files	109
5.5	R-base importing functions	110
5.5.1	<code>scan</code>	110
5.6	Text versus binary files	111
5.7	Unicode versus ASCII	111
5.8	Organizing data with spreadsheets	112
5.9	Exercises	112
II	Data Visualization	113
6	Introduction to data visualization	115

<i>0.0 Contents</i>	7
7 ggplot2	119
7.1 The components of a graph	120
7.2 ggplot objects	121
7.3 Geometries	122
7.4 Aesthetic mappings	123
7.5 Layers	124
7.5.1 Tinkering with arguments	125
7.6 Global versus local aesthetic mappings	126
7.7 Scales	127
7.8 Labels and titles	128
7.9 Categories as colors	129
7.10 Annotation, shapes, and adjustments	130
7.11 Add-on packages	131
7.12 Putting it all together	132
7.13 Quick plots with <code>qplot</code>	133
7.14 Grids of plots	134
7.15 Exercises	134
8 Visualizing data distributions	137
8.1 Variable types	137
8.2 Case study: describing student heights	138
8.3 Distribution function	138
8.4 Cumulative distribution functions	139
8.5 Histograms	140
8.6 Smoothed density	141
8.6.1 Interpreting the y-axis	145
8.6.2 Densities permit stratification	146
8.7 Exercises	146
8.8 The normal distribution	150
8.9 Standard units	152
8.10 Quantile-quantile plots	153
8.11 Percentiles	155
8.12 Boxplots	155
8.13 Stratification	157
8.14 Case study: describing student heights (continued)	157

8.15	Exercises	159
8.16	ggplot2 geometries	160
8.16.1	Barplots	161
8.16.2	Histograms	162
8.16.3	Density plots	163
8.16.4	Boxplots	164
8.16.5	QQ-plots	164
8.16.6	Images	165
8.16.7	Quick plots	166
8.17	Exercises	168
9	Data visualization in practice	169
9.1	Case study: new insights on poverty	169
9.1.1	Hans Rosling's quiz	170
9.2	Scatterplots	171
9.3	Faceting	172
9.3.1	<code>facet_wrap</code>	174
9.3.2	Fixed scales for better comparisons	175
9.4	Time series plots	175
9.4.1	Labels instead of legends	178
9.5	Data transformations	179
9.5.1	Log transformation	179
9.5.2	Which base?	181
9.5.3	Transform the values or the scale?	182
9.6	Visualizing multimodal distributions	183
9.7	Comparing multiple distributions with boxplots and ridge plots	183
9.7.1	Boxplots	184
9.7.2	Ridge plots	185
9.7.3	Example: 1970 versus 2010 income distributions	187
9.7.4	Accessing computed variables	193
9.7.5	Weighted densities	196
9.8	The ecological fallacy and importance of showing the data	196
9.8.1	Logistic transformation	197
9.8.2	Show the data	197

<i>0.0 Contents</i>	9
10 Data visualization principles	199
10.1 Encoding data using visual cues	199
10.2 Know when to include 0	202
10.3 Do not distort quantities	205
10.4 Order categories by a meaningful value	207
10.5 Show the data	208
10.6 Ease comparisons	211
10.6.1 Use common axes	211
10.6.2 Align plots vertically to see horizontal changes and horizontally to see vertical changes	212
10.6.3 Consider transformations	213
10.6.4 Visual cues to be compared should be adjacent	215
10.6.5 Use color	216
10.7 Think of the color blind	216
10.8 Plots for two variables	217
10.8.1 Slope charts	217
10.8.2 Bland-Altman plot	219
10.9 Encoding a third variable	219
10.10 Avoid pseudo-three-dimensional plots	221
10.11 Avoid too many significant digits	223
10.12 Know your audience	224
10.13 Exercises	224
10.14 Case study: vaccines and infectious diseases	229
10.15 Exercises	232
11 Robust summaries	233
11.1 Outliers	233
11.2 Median	234
11.3 The inter quartile range (IQR)	234
11.4 Tukey's definition of an outlier	235
11.5 Median absolute deviation	236
11.6 Exercises	236
11.7 Case study: self-reported student heights	237
III Statistics with R	241

12 Introduction to statistics with R	243
13 Probability	245
13.1 Discrete probability	245
13.1.1 Relative frequency	245
13.1.2 Notation	246
13.1.3 Probability distributions	246
13.2 Monte Carlo simulations for categorical data	246
13.2.1 Setting the random seed	248
13.2.2 With and without replacement	248
13.3 Independence	249
13.4 Conditional probabilities	249
13.5 Addition and multiplication rules	250
13.5.1 Multiplication rule	250
13.5.2 Multiplication rule under independence	250
13.5.3 Addition rule	251
13.6 Combinations and permutations	251
13.6.1 Monte Carlo example	255
13.7 Examples	255
13.7.1 Monty Hall problem	256
13.7.2 Birthday problem	257
13.8 Infinity in practice	259
13.9 Exercises	260
13.10 Continuous probability	262
13.11 Theoretical continuous distributions	263
13.11.1 Theoretical distributions as approximations	263
13.11.2 The probability density	265
13.12 Monte Carlo simulations for continuous variables	266
13.13 Continuous distributions	267
13.14 Exercises	267
14 Random variables	269
14.1 Random variables	269
14.2 Sampling models	270
14.3 The probability distribution of a random variable	271

<i>0.0 Contents</i>	11
14.4 Distributions versus probability distributions	273
14.5 Notation for random variables	273
14.6 The expected value and standard error	274
14.6.1 Population SD versus the sample SD	276
14.7 Central Limit Theorem	277
14.7.1 How large is large in the Central Limit Theorem?	278
14.8 Statistical properties of averages	278
14.9 Law of large numbers	280
14.9.1 Misinterpreting law of averages	280
14.10 Exercises	280
14.11 Case study: The Big Short	282
14.11.1 Interest rates explained with chance model	282
14.11.2 The Big Short	285
14.12 Exercises	288
15 Statistical inference	289
15.1 Polls	289
15.1.1 The sampling model for polls	290
15.2 Populations, samples, parameters, and estimates	292
15.2.1 The sample average	292
15.2.2 Parameters	293
15.2.3 Polling versus forecasting	293
15.2.4 Properties of our estimate: expected value and standard error	294
15.3 Exercises	295
15.4 Central Limit Theorem in practice	296
15.4.1 A Monte Carlo simulation	297
15.4.2 The spread	299
15.4.3 Bias: why not run a very large poll?	299
15.5 Exercises	300
15.6 Confidence intervals	302
15.6.1 A Monte Carlo simulation	304
15.6.2 The correct language	305
15.7 Exercises	305
15.8 Power	306
15.9 p-values	307

15.10 Association tests	308
15.10.1 Lady Tasting Tea	309
15.10.2 Two-by-two tables	310
15.10.3 Chi-square Test	310
15.10.4 The odds ratio	311
15.10.5 Confidence intervals for the odds ratio	312
15.10.6 Small count correction	313
15.10.7 Large samples, small p-values	313
15.11 Exercises	314
16 Statistical models	315
16.1 Poll aggregators	316
16.1.1 Poll data	318
16.1.2 Pollster bias	320
16.2 Data-driven models	321
16.3 Exercises	323
16.4 Bayesian statistics	326
16.4.1 Bayes theorem	326
16.5 Bayes theorem simulation	327
16.5.1 Bayes in practice	328
16.6 Hierarchical models	329
16.7 Exercises	331
16.8 Case study: election forecasting	333
16.8.1 Bayesian approach	334
16.8.2 The general bias	335
16.8.3 Mathematical representations of models	335
16.8.4 Predicting the electoral college	338
16.8.5 Forecasting	342
16.9 Exercises	345
16.10 The t-distribution	346
17 Regression	349
17.1 Case study: is height hereditary?	349
17.2 The correlation coefficient	350
17.2.1 Sample correlation is a random variable	352

0.0	Contents	13
17.2.2	Correlation is not always a useful summary	354
17.3	Conditional expectations	354
17.4	The regression line	357
17.4.1	Regression improves precision	358
17.4.2	Bivariate normal distribution (advanced)	359
17.4.3	Variance explained	361
17.4.4	Warning: there are two regression lines	361
17.5	Exercises	362
18	Linear models	363
18.1	Case study: Moneyball	363
18.1.1	Sabermetrics	364
18.1.2	Baseball basics	365
18.1.3	No awards for BB	366
18.1.4	Base on balls or stolen bases?	367
18.1.5	Regression applied to baseball statistics	369
18.2	Confounding	372
18.2.1	Understanding confounding through stratification	373
18.2.2	Multivariate regression	376
18.3	Least squares estimates	376
18.3.1	Interpreting linear models	377
18.3.2	Least Squares Estimates (LSE)	377
18.3.3	The <code>lm</code> function	379
18.3.4	LSE are random variables	380
18.3.5	Predicted values are random variables	381
18.4	Exercises	382
18.5	Linear regression in the tidyverse	383
18.5.1	The broom package	386
18.6	Exercises	387
18.7	Case study: Moneyball (continued)	388
18.7.1	Adding salary and position information	392
18.7.2	Picking nine players	393
18.8	The regression fallacy	395
18.9	Measurement error models	396
18.10	Exercises	399

19 Association is not causation	401
19.1 Spurious correlation	401
19.2 Outliers	404
19.3 Reversing cause and effect	406
19.4 Confounders	407
19.4.1 Example: UC Berkeley admissions	407
19.4.2 Confounding explained graphically	408
19.4.3 Average after stratifying	409
19.5 Simpson's paradox	410
19.6 Exercises	411
IV Data Wrangling	413
20 Introduction to data wrangling	415
21 Reshaping data	417
21.1 gather	417
21.2 spread	419
21.3 separate	419
21.4 unite	422
21.5 Exercises	423
22 Joining tables	425
22.1 Joins	426
22.1.1 Left join	427
22.1.2 Right join	428
22.1.3 Inner join	428
22.1.4 Full join	428
22.1.5 Semi join	429
22.1.6 Anti join	429
22.2 Binding	430
22.2.1 Binding columns	430
22.2.2 Binding by rows	430
22.3 Set operators	431
22.3.1 Intersect	431
22.3.2 Union	432

<i>0.0 Contents</i>	15
22.3.3 <code>setdiff</code>	432
22.3.4 <code>setequal</code>	432
22.4 Exercises	433
23 Web scraping	435
23.1 HTML	436
23.2 The <code>rvest</code> package	437
23.3 CSS selectors	439
23.4 JSON	440
23.5 Exercises	441
24 String processing	443
24.1 The <code>stringr</code> package	443
24.2 Case study 1: US murders data	445
24.3 Case study 2: self-reported heights	447
24.4 How to <i>escape</i> when defining strings	449
24.5 Regular expressions	451
24.5.1 Strings are a regexp	451
24.5.2 Special characters	451
24.5.3 Character classes	453
24.5.4 Anchors	454
24.5.5 Quantifiers	454
24.5.6 White space <code>\s</code>	455
24.5.7 Quantifiers: <code>*</code> , <code>?</code> , <code>+</code>	456
24.5.8 Not	456
24.5.9 Groups	457
24.6 Search and replace with regex	458
24.6.1 Search and replace using groups	460
24.7 Testing and improving	461
24.8 Trimming	463
24.9 Changing lettercase	464
24.10 Case study 2: self-reported heights (continued)	464
24.10.1 The <code>extract</code> function	465
24.10.2 Putting it all together	466
24.11 String splitting	467

24.12	Case study 3: extracting tables from a PDF	470
24.13	Recoding	473
24.14	Exercises	474
25	Parsing dates and times	477
25.1	The date data type	477
25.2	The lubridate package	478
25.3	Exercises	481
26	Text mining	483
26.1	Case study: Trump tweets	483
26.2	Text as data	485
26.3	Sentiment analysis	490
26.4	Exercises	495
V	Machine Learning	497
27	Introduction to machine learning	499
27.1	Notation	499
27.2	An example	500
27.3	Exercises	502
27.4	Evaluation metrics	502
27.4.1	Training and test sets	503
27.4.2	Overall accuracy	504
27.4.3	The confusion matrix	506
27.4.4	Sensitivity and specificity	507
27.4.5	Balanced accuracy and F_1 score	509
27.4.6	Prevalence matters in practice	510
27.4.7	ROC and precision-recall curves	511
27.4.8	The loss function	512
27.5	Exercises	514
27.6	Conditional probabilities and expectations	514
27.6.1	Conditional probabilities	515
27.6.2	Conditional expectations	516
27.6.3	Conditional expectation minimizes squared loss function	516
27.7	Exercises	517
27.8	Case study: is it a 2 or a 7?	517

<i>0.0 Contents</i>	17
28 Smoothing	521
28.1 Bin smoothing	523
28.2 Kernels	525
28.3 Local weighted regression (loess)	526
28.3.1 Fitting parabolas	530
28.3.2 Beware of default smoothing parameters	531
28.4 Connecting smoothing to machine learning	532
28.5 Exercises	532
29 Cross validation	535
29.1 Motivation with k-nearest neighbors	535
29.1.1 Over-training	537
29.1.2 Over-smoothing	538
29.1.3 Picking the k in kNN	539
29.2 Mathematical description of cross validation	541
29.3 K-fold cross validation	542
29.4 Exercises	545
29.5 Bootstrap	546
29.6 Exercises	549
30 The caret package	551
30.1 The caret <code>train</code> function	551
30.2 Cross validation	552
30.3 Example: fitting with loess	554
31 Examples of algorithms	557
31.1 Linear regression	557
31.1.1 The <code>predict</code> function	558
31.2 Exercises	559
31.3 Logistic regression	561
31.3.1 Generalized linear models	562
31.3.2 Logistic regression with more than one predictor	566
31.4 Exercises	567
31.5 k-nearest neighbors	568
31.6 Exercises	569
31.7 Generative models	569

31.7.1 Naive Bayes	570
31.7.2 Controlling prevalence	571
31.7.3 Quadratic discriminant analysis	573
31.7.4 Linear discriminant analysis	575
31.7.5 Connection to distance	577
31.8 Case study: more than three classes	577
31.9 Exercises	581
31.10 Classification and regression trees (CART)	582
31.10.1 The curse of dimensionality	582
31.10.2 CART motivation	583
31.10.3 Regression trees	586
31.10.4 Classification (decision) trees	592
31.11 Random forests	594
31.12 Exercises	599
32 Machine learning in practice	601
32.1 Preprocessing	602
32.2 k-nearest neighbor and random forest	603
32.3 Variable importance	606
32.4 Visual assessments	607
32.5 Ensembles	607
32.6 Exercises	608
33 Large datasets	609
33.1 Matrix algebra	609
33.1.1 Notation	610
33.1.2 Converting a vector to a matrix	612
33.1.3 Row and column summaries	613
33.1.4 <code>apply</code>	614
33.1.5 Filtering columns based on summaries	614
33.1.6 Indexing with matrices	616
33.1.7 Binarizing the data	618
33.1.8 Vectorization for matrices	618
33.1.9 Matrix algebra operations	619
33.2 Exercises	619

0.0 Contents	19
33.3 Distance	619
33.3.1 Euclidean distance	620
33.3.2 Distance in higher dimensions	620
33.3.3 Euclidean distance example	621
33.3.4 Predictor space	623
33.3.5 Distance between predictors	623
33.4 Exercises	623
33.5 Dimension reduction	624
33.5.1 Preserving distance	624
33.5.2 Linear transformations (advanced)	627
33.5.3 Orthogonal transformations (advanced)	628
33.5.4 Principal component analysis	630
33.5.5 Iris example	632
33.5.6 MNIST example	635
33.6 Exercises	637
33.7 Recommendation systems	638
33.7.1 Movielens data	638
33.7.2 Recommendation systems as a machine learning challenge	640
33.7.3 Loss function	640
33.7.4 A first model	641
33.7.5 Modeling movie effects	642
33.7.6 User effects	643
33.8 Exercises	644
33.9 Regularization	645
33.9.1 Motivation	645
33.9.2 Penalized least squares	647
33.9.3 Choosing the penalty terms	650
33.10 Exercises	652
33.11 Matrix factorization	653
33.11.1 Factors analysis	656
33.11.2 Connection to SVD and PCA	658
33.12 Exercises	661

34 Clustering	667
34.1 Hierarchical clustering	668
34.2 k-means	670
34.3 Heatmaps	670
34.4 Filtering features	671
34.5 Exercises	672
 VI Productivity Tools	 673
 35 Introduction to productivity tools	 675
 36 Organizing with Unix	 677
36.1 Naming convention	677
36.2 The terminal	678
36.3 The filesystem	678
36.3.1 Directories and subdirectories	679
36.3.2 The home directory	679
36.3.3 Working directory	680
36.3.4 Paths	681
36.4 Unix commands	681
36.4.1 <code>ls</code> : Listing directory content	682
36.4.2 <code>mkdir</code> and <code>rmdir</code> : make and remove a directory	682
36.4.3 <code>cd</code> : navigating the filesystem by changing directories	683
36.5 Some examples	685
36.6 More Unix commands	686
36.6.1 <code>mv</code> : moving files	686
36.6.2 <code>cp</code> : copying files	687
36.6.3 <code>rm</code> : removing files	687
36.6.4 <code>less</code> : looking at a file	688
36.7 Preparing for a data science project	688
36.8 Advanced Unix	689
36.8.1 Arguments	689
36.8.2 Getting help	690
36.8.3 Pipes	691
36.8.4 Wild cards	691
36.8.5 Environment variables	692

<i>0.0 Contents</i>	21
36.8.6 Shells	692
36.8.7 Executables	693
36.8.8 Permissions and file types	693
36.8.9 Commands you should learn	694
36.8.10 File manipulation in R	694
37 Git and GitHub	695
37.1 Why use Git and GitHub?	695
37.2 GitHub accounts	695
37.3 GitHub repositories	698
37.4 Overview of Git	699
37.4.1 Clone	700
37.5 Initializing a Git directory	704
37.6 Using Git and GitHub in RStudio	706
38 Reproducible projects with RStudio and R markdown	711
38.1 RStudio projects	711
38.2 R markdown	714
38.2.1 The header	716
38.2.2 R code chunks	716
38.2.3 Global options	717
38.2.4 knitR	717
38.2.5 More on R markdown	718
38.3 Organizing a data science project	718
38.3.1 Create directories in Unix	718
38.3.2 Create an RStudio project	719
38.3.3 Edit some R scripts	720
38.3.4 Create some more directories using Unix	721
38.3.5 Add a README file	721
38.3.6 Initializing a Git directory	721

