

A Hands-On Introduction to Data Science

CHIRAG SHAH

University of Washington



CAMBRIDGE
UNIVERSITY PRESS

Contents

<i>Preface</i>	<i>page xv</i>
<i>About the Author</i>	<i>xx</i>
<i>Acknowledgments</i>	<i>xxii</i>

Part I: Conceptual Introductions 1

1 Introduction	3
1.1 What Is Data Science?	3
1.2 Where Do We See Data Science?	5
1.2.1 Finance	6
1.2.2 Public Policy	7
1.2.3 Politics	8
1.2.4 Healthcare	9
1.2.5 Urban Planning	10
1.2.6 Education	10
1.2.7 Libraries	11
1.3 How Does Data Science Relate to Other Fields?	11
1.3.1 Data Science and Statistics	12
1.3.2 Data Science and Computer Science	13
1.3.3 Data Science and Engineering	13
1.3.4 Data Science and Business Analytics	14
1.3.5 Data Science, Social Science, and Computational Social Science	14
1.4 The Relationship between Data Science and Information Science	15
1.4.1 Information vs. Data	16
1.4.2 Users in Information Science	16
1.4.3 Data Science in Information Schools (iSchools)	17
1.5 Computational Thinking	17
1.6 Skills for Data Science	21
1.7 Tools for Data Science	27
1.8 Issues of Ethics, Bias, and Privacy in Data Science	29
Summary	30
Key Terms	31
Conceptual Questions	32
Hands-On Problems	32

2 Data	37
2.1 Introduction	37
2.2 Data Types	37
2.2.1 Structured Data	38
2.2.2 Unstructured Data	38
2.2.3 Challenges with Unstructured Data	39
2.3 Data Collections	39
2.3.1 Open Data	40
2.3.2 Social Media Data	41
2.3.3 Multimodal Data	41
2.3.4 Data Storage and Presentation	42
2.4 Data Pre-processing	47
2.4.1 Data Cleaning	48
2.4.2 Data Integration	50
2.4.3 Data Transformation	51
2.4.4 Data Reduction	51
2.4.5 Data Discretization	52
Summary	59
Key Terms	60
Conceptual Questions	60
Hands-On Problems	61
Further Reading and Resources	65
3 Techniques	66
3.1 Introduction	66
3.2 Data Analysis and Data Analytics	67
3.3 Descriptive Analysis	67
3.3.1 Variables	68
3.3.2 Frequency Distribution	71
3.3.3 Measures of Centrality	75
3.3.4 Dispersion of a Distribution	77
3.4 Diagnostic Analytics	82
3.4.1 Correlations	82
3.5 Predictive Analytics	84
3.6 Prescriptive Analytics	85
3.7 Exploratory Analysis	86
3.8 Mechanistic Analysis	87
3.8.1 Regression	87
Summary	89
Key Terms	91
Conceptual Questions	92
Hands-On Problems	92
Further Reading and Resources	95

Part II: Tools for Data Science	97
4 UNIX	99
4.1 Introduction	99
4.2 Getting Access to UNIX	100
4.3 Connecting to a UNIX Server	102
4.3.1 SSH	102
4.3.2 FTP/SCP/SFTP	104
4.4 Basic Commands	106
4.4.1 File and Directory Manipulation Commands	106
4.4.2 Process-Related Commands	108
4.4.3 Other Useful Commands	109
4.4.4 Shortcuts	109
4.5 Editing on UNIX	110
4.5.1 The vi Editor	110
4.5.2 The Emacs Editor	111
4.6 Redirections and Piping	112
4.7 Solving Small Problems with UNIX	113
Summary	121
Key Terms	121
Conceptual Questions	122
Hands-On Problems	122
Further Reading and Resources	123
5 Python	125
5.1 Introduction	125
5.2 Getting Access to Python	125
5.2.1 Download and Install Python	126
5.2.2 Running Python through Console	126
5.2.3 Using Python through Integrated Development Environment (IDE)	126
5.3 Basic Examples	128
5.4 Control Structures	131
5.5 Statistics Essentials	133
5.5.1 Importing Data	136
5.5.2 Plotting the Data	137
5.5.3 Correlation	138
5.5.4 Linear Regression	138
5.5.5 Multiple Linear Regression	141
5.6 Introduction to Machine Learning	145
5.6.1 What Is Machine Learning?	145
5.6.2 Classification (Supervised Learning)	147
5.6.3 Clustering (Unsupervised Learning)	150
5.6.4 Density Estimation (Unsupervised Learning)	153

Summary	155
Key Terms	156
Conceptual Questions	157
Hands-On Problems	157
Further Reading and Resources	159
6 R	161
6.1 Introduction	161
6.2 Getting Access to R	162
6.3 Getting Started with R	163
6.3.1 Basics	163
6.3.2 Control Structures	165
6.3.3 Functions	167
6.3.4 Importing Data	167
6.4 Graphics and Data Visualization	168
6.4.1 Installing ggplot2	168
6.4.2 Loading the Data	169
6.4.3 Plotting the Data	169
6.5 Statistics and Machine Learning	174
6.5.1 Basic Statistics	174
6.5.2 Regression	176
6.5.3 Classification	178
6.5.4 Clustering	180
Summary	182
Key Terms	183
Conceptual Questions	184
Hands-On Problems	184
Further Reading and Resources	185
7 MySQL	187
7.1 Introduction	187
7.2 Getting Started with MySQL	188
7.2.1 Obtaining MySQL	188
7.2.2 Logging in to MySQL	188
7.3 Creating and Inserting Records	191
7.3.1 Importing Data	191
7.3.2 Creating a Table	192
7.3.3 Inserting Records	192
7.4 Retrieving Records	193
7.4.1 Reading Details about Tables	193
7.4.2 Retrieving Information from Tables	193
7.5 Searching in MySQL	195
7.5.1 Searching within Field Values	195
7.5.2 Full-Text Searching with Indexing	195

7.6	Accessing MySQL with Python	196
7.7	Accessing MySQL with R	199
7.8	Introduction to Other Popular Databases	200
7.8.1	NoSQL	200
7.8.2	MongoDB	201
7.8.3	Google BigQuery	201
	Summary	202
	Key Terms	202
	Conceptual Questions	203
	Hands-On Problems	203
	Further Reading and Resources	204
	Part III: Machine Learning for Data Science	207
8	Machine Learning Introduction and Regression	209
8.1	Introduction	209
8.2	What Is Machine Learning?	210
8.3	Regression	215
8.4	Gradient Descent	220
	Summary	229
	Key Terms	230
	Conceptual Questions	231
	Hands-On Problems	231
	Further Reading and Resources	233
9	Supervised Learning	235
9.1	Introduction	235
9.2	Logistic Regression	236
9.3	Softmax Regression	244
9.4	Classification with kNN	248
9.5	Decision Tree	252
9.5.1	Decision Rule	256
9.5.2	Classification Rule	257
9.5.3	Association Rule	257
9.6	Random Forest	260
9.7	Naïve Bayes	266
9.8	Support Vector Machine (SVM)	272
	Summary	279
	Key Terms	280
	Conceptual Questions	281
	Hands-On Problems	281
	Further Reading and Resources	288

10 Unsupervised Learning	290
10.1 Introduction	290
10.2 Agglomerative Clustering	291
10.3 Divisive Clustering	295
10.4 Expectation Maximization (EM)	299
10.5 Introduction to Reinforcement Learning	309
Summary	312
Key Terms	313
Conceptual Questions	314
Hands-On Problems	314
Further Reading and Resources	317
Part IV: Applications, Evaluations, and Methods	319
11 Hands-On with Solving Data Problems	321
11.1 Introduction	321
11.2 Collecting and Analyzing Twitter Data	328
11.3 Collecting and Analyzing YouTube Data	336
11.4 Analyzing Yelp Reviews and Ratings	342
Summary	349
Key Terms	350
Conceptual Questions	350
Practice Questions	351
12 Data Collection, Experimentation, and Evaluation	354
12.1 Introduction	354
12.2 Data Collection Methods	355
12.2.1 Surveys	355
12.2.2 Survey Question Types	355
12.2.3 Survey Audience	357
12.2.4 Survey Services	358
12.2.5 Analyzing Survey Data	359
12.2.6 Pros and Cons of Surveys	360
12.2.7 Interviews and Focus Groups	360
12.2.8 Why Do an Interview?	360
12.2.9 Why Focus Groups?	361
12.2.10 Interview or Focus Group Procedure	361
12.2.11 Analyzing Interview Data	362
12.2.12 Pros and Cons of Interviews and Focus Groups	362
12.2.13 Log and Diary Data	363
12.2.14 User Studies in Lab and Field	364
12.3 Picking Data Collection and Analysis Methods	366
12.3.1 Introduction to Quantitative Methods	366

12.3.2	Introduction to Qualitative Methods	368
12.3.3	Mixed Method Studies	369
12.4	Evaluation	370
12.4.1	Comparing Models	370
12.4.2	Training–Testing and A/B Testing	372
12.4.3	Cross-Validation	374
	Summary	376
	Key Terms	377
	Conceptual Questions	377
	Further Reading and Resources	378
 <i>Appendices</i>		
	<i>Appendix A: Useful Formulas from Differential Calculus</i>	379
	Further Reading and Resources	380
	<i>Appendix B: Useful Formulas from Probability</i>	381
	Further Reading and Resources	381
	<i>Appendix C: Useful Resources</i>	383
	C.1 Tutorials	383
	C.2 Tools	383
	<i>Appendix D: Installing and Configuring Tools</i>	385
	D.1 Anaconda	385
	D.2 IPython (Jupyter) Notebook	385
	D.3 Spyder	387
	D.4 R	387
	D.5 RStudio	388
	<i>Appendix E: Datasets and Data Challenges</i>	390
	E.1 Kaggle	390
	E.2 RecSys	391
	E.3 WSDM	391
	E.4 KDD Cup	392
	<i>Appendix F: Using Cloud Services</i>	393
	F.1 Google Cloud Platform	394
	F.2 Hadoop	398
	F.3 Microsoft Azure	400
	F.4 Amazon Web Services (AWS)	403
	<i>Appendix G: Data Science Jobs</i>	407
	G.1 Marketing	408
	G.2 Corporate Retail and Sales	409
	G.3 Legal	409
	G.4 Health and Human Services	410

<i>Appendix H: Data Science and Ethics</i>	412
H.1 Data Supply Chain	412
H.2 Bias and Inclusion	414
H.3 Considering Best Practices and Codes of Conduct	414
<i>Appendix I: Data Science for Social Good</i>	416
<i>Index</i>	418