

Contents

<i>Preface</i>	<i>page xi</i>
<i>About the Author</i>	<i>xvii</i>
Part I Conceptual Introductions	1
1 Introduction	3
Online Datasets	3
1.1 What Is Data Science?	3
1.2 Where Do We See Data Science?	6
1.2.1 Finance	6
1.2.2 Public Policy	7
1.2.3 Politics	8
1.2.4 Healthcare	9
1.2.5 Climate Change	10
1.2.6 Urban Planning	10
1.2.7 Education	11
1.2.8 Libraries	12
1.3 How Does Data Science Relate to Other Fields?	13
1.3.1 Data Science and Statistics	13
1.3.2 Data Science and Computer Science	14
1.3.3 Data Science and Engineering	14
1.3.4 Data Science and Business Analytics	15
1.3.5 Data Science, Social Science, and Computational Social Science	15
1.4 The Relationship between Data Science and Information Science	16
1.4.1 Information vs. Data	17
1.4.2 Users in Information Science	17
1.4.3 Data Science in Information Schools (iSchools)	18
1.5 The Relationship between Data Science and Artificial Intelligence	18
1.6 Computational Thinking	20
1.7 Skills for Data Science	23
1.8 Tools for Data Science	29
1.9 Issues of Ethics, Bias, and Privacy in Data Science	30
1.9.1 Concerns for Data Users	31
1.9.2 Concerns for Data Scientists	32
1.9.3 Data Supply Chain	33
1.9.4 Bias and Inclusion	34
1.9.5 Considering Best Practices and Codes of Conduct	34
Summary	35

Key Terms	36
Conceptual Questions	37
Hands-On Problems	37
References	40
2 Data	43
Online Datasets	43
2.1 Introduction	43
2.2 Data Types	44
2.2.1 Structured Data	44
2.2.2 Unstructured Data	45
2.2.3 Challenges with Unstructured Data	45
2.3 Data Collections	46
2.3.1 Open Data	46
2.3.2 Social Media Data	47
2.3.3 Multimodal Data	47
2.3.4 Synthetic Data	48
2.4 Data Storage and Presentation	49
2.5 Data Pre-Processing	54
2.5.1 Data Cleaning	55
2.5.2 Data Integration	57
2.5.3 Data Transformation	58
2.5.4 Data Reduction	58
2.5.5 Data Discretization	59
Summary	67
Key Terms	67
Conceptual Questions	68
Hands-On Problems	68
Further Reading and Resources	73
References	73
3 Techniques	75
Online Appendix	75
3.1 Introduction	75
3.2 Data Analysis and Data Analytics	76
3.3 Descriptive Analysis	77
3.3.1 Variables	78
3.3.2 Frequency Distribution	81
3.3.3 Measures of Centrality	85
3.3.4 Dispersion of a Distribution	87
3.4 Diagnostic Analytics	91
3.4.1 Correlations	92
3.5 Predictive Analytics	95
3.6 Prescriptive Analytics	96
3.7 Exploratory Analysis	97
3.8 Mechanistic Analysis	98
3.8.1 Regression	98

Summary	100
Key Terms	102
Conceptual Questions	103
Hands-On Problems	104
Further Reading and Resources	107
References	107
Part II Tools for Data Science	
4 Python	111
4.1 Introduction	111
4.2 Getting Access to Python	111
4.2.1 Download and Install Python	112
4.2.2 Running Python through Console	112
4.2.3 Using Python through Integrated Development Environment (IDE)	112
4.3 Basic Examples	115
4.4 Data Structures	117
4.5 Control Structures	120
4.6 Functions	122
4.7 Making Python Interactive	124
4.8 Installing and Using Python Packages	125
Summary	127
Key Terms	128
Conceptual Questions	128
Hands-On Problems	129
Further Reading and Resources	130
References	131
5 Python for Statistical Analysis	132
Online Datasets	132
5.1 Introduction	132
5.2 Statistics Essentials	133
5.3 Graphics and Data Visualization	136
5.3.1 Importing Data	136
5.3.2 Plotting the Data	136
5.4 Statistical Inference	137
5.4.1 Correlation	138
5.4.2 Hypothesis Testing	139
5.4.3 Comparing Means Using t-Test	140
5.4.4 Analysis of Variance Using ANOVA	143
Summary	145
Key Terms	146
Conceptual Questions	146
Hands-On Problems	147
Further Reading and Resources	149

6 Cloud Computing	151
6.1 Cloud Computing	151
6.2 Google Cloud Platform	152
6.2.1 Hadoop	155
6.3 Microsoft Azure	161
6.4 Amazon Web Services (AWS)	169
6.5 Moving between Cloud Platforms	176
Summary	177
Key Terms	177
Conceptual Questions	178
Hands-on Problems	178
References	179
Part III Machine Learning for Data Science	181
7 Machine Learning Introduction and Regression	183
Online Datasets	183
7.1 Introduction	183
7.2 What Is Machine Learning?	184
7.3 Regression	189
7.4 Gradient Descent	195
7.5 Considerations for Applying Machine Learning Techniques	203
Summary	205
Key Terms	206
Conceptual Questions	206
Hands-On Problems	207
Further Reading and Resources	208
References	209
8 Supervised Learning	210
Online Datasets	210
8.1 Introduction	211
8.2 Logistic Regression	212
8.3 Softmax Regression	221
8.4 Classification with kNN	225
8.5 Decision Tree	230
8.5.1 Decision Rule	235
8.5.2 Classification Rule	236
8.5.3 Association Rule	236
8.6 Random Forest	239
8.7 Naive Bayes	245
8.8 Support Vector Machine (SVM)	252
Summary	260
Key Terms	261
Conceptual Questions	262
Hands-On Problems	262

Further Reading and Resources	268
References	269
9 Unsupervised Learning	270
Online Datasets	270
9.1 Introduction	270
9.2 Agglomerative Clustering	271
9.3 Divisive Clustering	275
9.4 Expectation Maximization (EM)	279
9.5 Introduction to Reinforcement Learning	289
Summary	294
Key Terms	295
Conceptual Questions	296
Hands-On Problems	296
Further Reading and Resources	298
References	299
Part IV Applications, Evaluations, and Methods	301
10 Data Collection, Experimentation, and Evaluation	303
10.1 Introduction	303
10.2 Data Collection Methods	304
10.2.1 Surveys	304
10.2.2 Survey Question Types	304
10.2.3 Survey Audience	306
10.2.4 Survey Services	307
10.2.5 Analyzing Survey Data	308
10.2.6 Pros and Cons of Surveys	308
10.2.7 Interviews and Focus Groups	309
10.2.8 Why Do an Interview?	309
10.2.9 Why Focus Groups?	310
10.2.10 Interview or Focus Group Procedure?	310
10.2.11 Analyzing Interview Data	311
10.2.12 Pros and Cons of Interviews and Focus Groups	312
10.2.13 Log and Diary Data	312
10.2.14 User Studies in Lab and Field	314
10.3 Picking Data Collection and Analysis Methods	315
10.3.1 Introduction to Quantitative Methods	316
10.3.2 Introduction to Qualitative Methods	317
10.3.3 Mixed Method Studies	318
10.4 Evaluation	319
10.4.1 Comparing Models	320
10.4.2 Training–Testing and A/B Testing	322
10.4.3 Cross-Validation	324
Summary	325
Key Terms	326

Conceptual Questions	327
Further Reading and Resources	327
References	327
11 Hands-On with Solving Data Problems	329
Online Datasets	329
11.1 Introduction	329
11.2 Collecting and Analyzing Reddit Data	336
11.3 Collecting and Analyzing YouTube Data	342
11.4 Analyzing Yelp Reviews and Ratings	349
Summary	355
Key Terms	356
Conceptual Questions	356
Hands-On Problems	356
References	359
<i>Appendix A: Useful Formulas</i>	360
<i>Appendix B: Installing and Configuring Tools</i>	363
B.1 Anaconda	363
B.2 IPython (Jupyter) Notebook	363
B.3 Spyder	363
<i>Appendix C: Using MySQL with Python</i>	366
C.1 Getting Started with MySQL	366
C.1.1 Obtaining MySQL	366
C.1.2 Logging in to MySQL	367
C.2 Creating and Inserting Records	369
C.2.1 Importing Data	369
C.2.2 Creating a Table	370
C.2.3 Inserting Records	371
C.3 Retrieving Records	371
C.3.1 Reading Details about Tables	372
C.3.2 Retrieving Information from Tables	372
C.4 Searching in MySQL	373
C.4.1 Searching within Field Values	374
C.4.2 Full-Text Searching with Indexing	374
C.5 Accessing MySQL with Python	375
<i>Appendix D: Introduction to Other Popular Databases</i>	378
D.1 NoSQL	378
D.2 MongoDB	378
D.3 Google BigQuery	379
<i>Appendix E: Data Science Jobs</i>	380
E.1 Marketing	381
E.2 Corporate Retail and Sales	381
E.3 Legal	382
E.4 Health and Human Services	383
<i>Index</i>	385