

---

# Table of Contents

<b>Preface.....</b>	<b>xi</b>
<b>1. Introduction.....</b>	<b>1</b>
The Ascendance of Data	1
What Is Data Science?	1
Motivating Hypothetical: DataSciencester	2
Finding Key Connectors	3
Data Scientists You May Know	6
Salaries and Experience	8
Paid Accounts	11
Topics of Interest	11
Onward	13
<b>2. A Crash Course in Python.....</b>	<b>15</b>
The Basics	15
Getting Python	15
The Zen of Python	16
Whitespace Formatting	16
Modules	17
Arithmetic	18
Functions	18
Strings	19
Exceptions	19
Lists	20
Tuples	21
Dictionaries	21
Sets	24
Control Flow	25

Truthiness	25
The Not-So-Basics	26
Sorting	27
List Comprehensions	27
Generators and Iterators	28
Randomness	29
Regular Expressions	30
Object-Oriented Programming	30
Functional Tools	31
enumerate	32
zip and Argument Unpacking	33
args and kwargs	34
Welcome to DataSciencester!	35
For Further Exploration	35
<b>3. Visualizing Data.....</b>	<b>37</b>
matplotlib	37
Bar Charts	39
Line Charts	43
Scatterplots	44
For Further Exploration	47
<b>4. Linear Algebra.....</b>	<b>49</b>
Vectors	49
Matrices	53
For Further Exploration	55
<b>5. Statistics.....</b>	<b>57</b>
Describing a Single Set of Data	57
Central Tendencies	59
Dispersion	61
Correlation	62
Simpson's Paradox	65
Some Other Correlational Caveats	66
Correlation and Causation	67
For Further Exploration	68
<b>6. Probability.....</b>	<b>69</b>
Dependence and Independence	69
Conditional Probability	70
Bayes's Theorem	72
Random Variables	73

Continuous Distributions	74
The Normal Distribution	75
The Central Limit Theorem	78
For Further Exploration	80
<b>7. Hypothesis and Inference.....</b>	<b>81</b>
Statistical Hypothesis Testing	81
Example: Flipping a Coin	81
Confidence Intervals	85
P-hacking	86
Example: Running an A/B Test	87
Bayesian Inference	88
For Further Exploration	92
<b>8. Gradient Descent.....</b>	<b>93</b>
The Idea Behind Gradient Descent	93
Estimating the Gradient	94
Using the Gradient	97
Choosing the Right Step Size	97
Putting It All Together	98
Stochastic Gradient Descent	99
For Further Exploration	100
<b>9. Getting Data.....</b>	<b>103</b>
stdin and stdout	103
Reading Files	105
The Basics of Text Files	105
Delimited Files	106
Scraping the Web	108
HTML and the Parsing Thereof	108
Example: O'Reilly Books About Data	110
Using APIs	114
JSON (and XML)	114
Using an Unauthenticated API	115
Finding APIs	116
Example: Using the Twitter APIs	117
Getting Credentials	117
For Further Exploration	120
<b>10. Working with Data.....</b>	<b>121</b>
Exploring Your Data	121
Exploring One-Dimensional Data	121

Two Dimensions	123
Many Dimensions	125
Cleaning and Munging	127
Manipulating Data	129
Rescaling	132
Dimensionality Reduction	134
For Further Exploration	139
<b>11. Machine Learning.....</b>	<b>141</b>
Modeling	141
What Is Machine Learning?	142
Overfitting and Underfitting	142
Correctness	145
The Bias-Variance Trade-off	147
Feature Extraction and Selection	148
For Further Exploration	150
<b>12. k-Nearest Neighbors.....</b>	<b>151</b>
The Model	151
Example: Favorite Languages	153
The Curse of Dimensionality	156
For Further Exploration	163
<b>13. Naive Bayes.....</b>	<b>165</b>
A Really Dumb Spam Filter	165
A More Sophisticated Spam Filter	166
Implementation	168
Testing Our Model	169
For Further Exploration	172
<b>14. Simple Linear Regression.....</b>	<b>173</b>
The Model	173
Using Gradient Descent	176
Maximum Likelihood Estimation	177
For Further Exploration	177
<b>15. Multiple Regression.....</b>	<b>179</b>
The Model	179
Further Assumptions of the Least Squares Model	180
Fitting the Model	181
Interpreting the Model	182
Goodness of Fit	183

Digression: The Bootstrap	183
Standard Errors of Regression Coefficients	184
Regularization	186
For Further Exploration	188
<b>16. Logistic Regression.....</b>	<b>189</b>
The Problem	189
The Logistic Function	192
Applying the Model	194
Goodness of Fit	195
Support Vector Machines	196
For Further Investigation	200
<b>17. Decision Trees.....</b>	<b>201</b>
What Is a Decision Tree?	201
Entropy	203
The Entropy of a Partition	205
Creating a Decision Tree	206
Putting It All Together	208
Random Forests	211
For Further Exploration	212
<b>18. Neural Networks.....</b>	<b>213</b>
Perceptrons	213
Feed-Forward Neural Networks	215
Backpropagation	218
Example: Defeating a CAPTCHA	219
For Further Exploration	224
<b>19. Clustering.....</b>	<b>225</b>
The Idea	225
The Model	226
Example: Meetups	227
Choosing k	230
Example: Clustering Colors	231
Bottom-up Hierarchical Clustering	233
For Further Exploration	238
<b>20. Natural Language Processing.....</b>	<b>239</b>
Word Clouds	239
n-gram Models	241
Grammars	244

An Aside: Gibbs Sampling	246
Topic Modeling	247
For Further Exploration	253
<b>21. Network Analysis.....</b>	<b>255</b>
Betweenness Centrality	255
Eigenvector Centrality	260
Matrix Multiplication	260
Centrality	262
Directed Graphs and PageRank	264
For Further Exploration	266
<b>22. Recommender Systems.....</b>	<b>267</b>
Manual Curation	268
Recommending What's Popular	268
User-Based Collaborative Filtering	269
Item-Based Collaborative Filtering	272
For Further Exploration	274
<b>23. Databases and SQL.....</b>	<b>275</b>
CREATE TABLE and INSERT	275
UPDATE	277
DELETE	278
SELECT	278
GROUP BY	280
ORDER BY	282
JOIN	283
Subqueries	285
Indexes	285
Query Optimization	286
NoSQL	287
For Further Exploration	287
<b>24. MapReduce.....</b>	<b>289</b>
Example: Word Count	289
Why MapReduce?	291
MapReduce More Generally	292
Example: Analyzing Status Updates	293
Example: Matrix Multiplication	294
An Aside: Combiners	296
For Further Exploration	296

<b>25. Go Forth and Do Data Science.....</b>	<b>299</b>
IPython	299
Mathematics	300
Not from Scratch	300
NumPy	301
pandas	301
scikit-learn	301
Visualization	301
R	302
Find Data	302
Do Data Science	303
Hacker News	303
Fire Trucks	303
T-shirts	304
And You?	304
<b>Index.....</b>	<b>305</b>