

Contents

Prerequisites: Tooling	1
Files Included with this Book	1
Python	2
Editor	5
Console (REPL)	5
Using Spyder and Keyboard Shortcuts	7
Anki	7
Remembering What You Learn	7
1. Introduction	8
The Purpose of Data Analysis	8
What is Data?	8
Example Datasets	10
100 Game Sample	10
2018 Season Data	12
What is Analysis?	14
Types of Data Analysis	15
Summary Statistics	15
Modeling	16
High Level Data Analysis Process	19
1. Collecting Data	19
2. Storing Data	19
3. Loading Data	19
4. Manipulating Data	19
5. Analyzing Data for Insights	20
Connecting the High Level Analysis Process to the Rest of the Book	20
End of Chapter Exercises	21
2. Python	23
Introduction to Python Programming	23
How to Read This Chapter	23

Important Parts of the Python Standard Library	24
Comments	24
Variables	25
Types	26
Interlude: How to Figure Things Out in Python	27
Bools	30
if statements	31
Container Types	32
Unpacking	34
Loops	34
Comprehensions	36
Functions	39
Libraries are Functions and Types	44
The os Library and path	44
End of Chapter Exercises	45
3. Pandas	48
Introduction to Pandas	48
Types and Functions	48
Things You Can Do with DataFrames	49
How to Read This Chapter	49
Part 1. DataFrame Basics	50
Loading Data	50
DataFrame Methods and Attributes	51
Working with Subsets of Columns	52
Indexing	53
Outputting Data	57
Exercises	58
Part 2. Things you can do with DataFrames	59
Introduction	59
1. Modify or create new columns of data	60
Modifying and creating columns are really the same thing	60
Math and number columns	61
String columns	62
Bool Columns	63
Applying Functions to Columns	64
Dropping Columns	65
Renaming Columns	65

Missing Data in Columns	67
Changing column types	69
Review	71
Exercises	72
2. Use built-in Pandas functions that work on DataFrames	73
Axis	74
Summary Functions on Boolean Columns	76
Other Misc Built-in Summary Functions	78
Review	79
Exercises	80
3. Filter Observations	81
loc	81
Boolean indexing	82
Duplicates	84
Combining filtering with changing columns	86
Query	87
Review	88
Exercises	89
4. Change Granularity	90
Ways of changing granularity	90
Grouping	90
A note on multilevel indexing	93
Stacking and unstacking data	94
Review	95
Exercises	96
5. Combining two or more DataFrames	97
1. The columns you're joining on.	97
2. Whether you're doing a one-to-one, one-to-many, or many-to-many merge	99
3. What you do with unmatched observations	100
More on pd.merge	102
pd.concat()	104
Review	106
Exercises	107
4. SQL	108
Introduction to SQL	108
How to Read This Chapter	108

Databases	108
SQL Databases	110
A Note on NoSQL	110
SQL	111
Pandas	111
Creating Data	111
Queries	112
Filtering	114
Joining, or Selecting From Multiple Tables	116
Misc SQL	121
SQL Example — LEFT JOIN, UNION, Subqueries	122
End of Chapter Exercises	126
5. Web Scraping and APIs	127
Introduction to Web Scraping and APIs	127
Web Scraping	127
HTML	128
BeautifulSoup	130
Baseball Almanac Opening Day History - Web Scraping Example	132
APIs	138
Two Types of APIs	138
HTTP	139
JSON	140
Benefits	141
MLB Data API Example	141
6. Distributions and Visualizing them in Python with Seaborn	159
Introduction	159
What is a Distribution?	160
Summary Stats: Info About Distributions	163
Introduction	163
Percentiles	163
Average = Expected Value	164
Dispersion	165
Density Plots in Python	165
Seaborn	165
Plot Options	174

7. Modeling	179
Introduction to Modeling	179
The Simplest Model	179
Linear regression	180
First linear regression: extended example	181
First linear regression: data processing	181
First linear regression: running the model	189
Statistical Significance	192
Regressions hold things constant	195
Fixed Effects	198
Squaring Variables	202
Logging Variables	204
Interactions	206
Logistic Regression	210
Random Forest	212
Classification and Regression Trees	212
Random Forests are a Bunch of Trees	213
Using a Trained Random Forest to Generate Predictions	213
Random Forest Example in Scikit-Learn	214
Random Forest Regressions	218
8. Intermediate Coding and Next Steps: High Level Strategies	220
Gall's Law	220
Get Quick Feedback	221
Use Functions	222
DRY: Don't Repeat Yourself	222
Functions Help You Think Less	222
Attitude	223
Review and Conclusion	225
Appendix A: Places to Get Data	226
Ready-made Datasets and APIs	226
Lahman's Baseball Data	226
Kaggle.com	226
Google Dataset Search	226
Appendix B: Anki	227
Remembering What You Learn	227
Installing Anki	228

Using Anki with this Book	229
Appendix C: Answers to End of Chapter Exercises	231
1. Introduction	231
2. Python	233
3.0 Pandas Basics	238
3.1 Columns	240
3.2 Built-in Functions	244
3.3 Filtering	246
3.4 Granularity	249
3.5 Combining DataFrames	252
4. SQL	254