

Master Data Analysis with Python

by
Ted Petrou

© 2019 Ted Petrou All Rights Reserved

Contents

I	Environment Setup and Jupyter Notebooks	1
1	Installing Python and Setting up an Environment for Data Science	3
1.1	Conda	3
1.2	Miniconda	3
1.3	Python and Conda installation complete	6
1.4	Creating a new environment just for data analysis	8
1.5	Create a new environment	9
1.6	Other Considerations	11
1.7	Summary of Steps	13
1.8	Using the book with Jupyter Notebooks	13
2	Introduction to Jupyter Notebooks	15
2.1	Jupyter Notebook Basics	15
2.2	Getting Started with Jupyter Notebooks	16
2.3	Executing Cells	17
2.4	Keyboard Shortcuts	18
2.5	Completing exercises in this book	20
2.6	Getting help in the notebook	20
3	Markdown Tutorial	23
3.1	Headers	23
3.2	Italics and bold	23
3.3	Code formatting	23
3.4	Lists	24
3.5	Hyperlinks	24
3.6	Images	25
3.7	New lines	25
4	Jupyter Notebooks More	27
4.1	Where Jupyter Notebooks Excel	27
4.2	Where Jupyter Notebooks Fail	27
4.3	On-demand Jupyter Notebooks	28
5	Jupyter Notebook Extensions	29
5.1	The NBextensions tab	29
5.2	The Skip-Traceback Extension	30
5.3	Skip-Traceback in the Notebook	30
II	Intro to pandas	33
6	What is pandas?	35
6.1	pandas operates on tabular (table) data	36
6.2	pandas examples	36
6.3	Reading data	36
6.4	Filtering data	37
6.5	Aggregating methods	38

6.6	Non-aggregating methods	39
6.7	Aggregating within groups	39
6.8	Tidying	42
6.9	Joining Data	44
6.10	Time Series Analysis	45
6.11	Visualization	46
6.12	Much More	48
7	The DataFrame and Series	49
7.1	Import pandas and read in data with <code>read_csv</code>	49
7.2	Components of a DataFrame	51
7.3	What type of object is <code>bikes</code> ?	52
7.4	Select a single column from a DataFrame - a Series	53
7.5	Components of a Series - the index and the data	53
7.6	Exercises	54
8	Data Types and Missing Values	57
8.1	Missing Value Representation	58
8.2	Finding the data types of each column	58
8.3	Getting more metadata	59
8.4	More data types	60
8.5	Exercises	61
9	Setting a Meaningful Index	63
9.1	Extracting the components of a DataFrame	63
9.2	Extracting the components of a Series	65
9.3	The default index	65
9.4	Setting an index on read	66
9.5	Choosing a good index	68
9.6	Setting the index with <code>set_index</code>	68
9.7	Changing Display Options	69
9.8	Exercises	69
10	Five-Step Process for Data Exploration	71
III	Selecting Subsets of Data	75
11	Selecting Subsets of Data from DataFrames with just the brackets	77
11.1	Selecting Subsets of Data	77
11.2	pandas dual references: by label and by integer location	78
11.3	The three indexers <code>[]</code> , <code>loc</code> , <code>iloc</code>	79
11.4	Begin with <i>just the brackets</i>	79
11.5	Select Multiple Columns with a List	80
11.6	Exercises	81
12	Selecting Subsets of Data from DataFrames with <code>loc</code>	83
12.1	Subset selection with <code>loc</code>	83
12.2	<code>loc</code> with slice notation	84
12.3	Select a single row and a single column	88
12.4	Summary of <code>loc</code>	88
12.5	Exercises	88
13	Selecting Subsets of Data from DataFrames with <code>iloc</code>	91
13.1	Getting started with <code>iloc</code>	91
13.2	Summary of <code>iloc</code>	95
13.3	Exercises	95
14	Selecting Subsets of Data from a Series	97
14.1	Using Dot Notation to Select a Column as a Series	97

14.2	Selecting Subsets of Data From a Series	98
14.3	Comparison to Python Lists and Dictionaries	100
14.4	Exercises	101
15	Boolean Selection Single Conditions	103
15.1	Boolean Selection	103
15.2	Manual filtering the data	104
15.3	Operator Overloading	105
15.4	Practical Boolean Selection	105
15.5	Boolean selection in one line	107
15.6	Single condition expression	108
15.7	Exercises	108
16	Boolean Selection Multiple Conditions	111
16.1	Multiple condition expression	111
16.2	Multiple conditions in one line	112
16.3	Using an or condition	113
16.4	Inverting a condition with the not operator	113
16.5	Lots of equality conditions in a single column - use <code>isin</code>	115
16.6	Exercises	117
17	Boolean Selection More	119
17.1	Boolean selection on a Series	119
17.2	The <code>between</code> method	120
17.3	Simultaneous boolean selection of rows and column labels with <code>loc</code>	121
17.4	Column to column comparisons	122
17.5	Finding Missing Values with <code>isna</code>	123
17.6	Exercises	124
18	Miscellaneous Subset Selection	127
18.1	The <code>query</code> method	127
18.2	Slicing with just the brackets	131
18.3	Selecting a single cell with <code>at</code> and <code>iat</code>	132
18.4	Exercises	133
IV	Essential Commands	135
19	Series Attributes and Statistical Methods	137
19.1	Calling methods on a Series/DataFrame	137
19.2	Begin with the Series	137
19.3	Core Series Attributes	138
19.4	Arithmetic operators	139
19.5	Comparison Operations	140
19.6	Boolean and bitwise operators	141
19.7	Statistical methods	141
19.8	Aggregation methods	141
19.9	Non-Aggregation methods	142
19.10	Operations on a boolean Series	143
19.11	Exercises	144
20	Series Methods More	147
20.1	Overview	147
20.2	Methods for handling missing values	147
20.3	Filling missing values	148
20.4	Dropping missing values	149
20.5	Sorting	149
20.6	Exercises	150

21 Series Methods More II	153
21.1 More accumulation methods	154
21.2 The rank method	155
21.3 Differencing methods <code>diff</code> and <code>pct_change</code>	155
21.4 Calling methods after an operation	156
21.5 Randomly sample a Series	157
21.6 Index of maximum and minimum	157
21.7 Uniqueness	158
21.8 Exercises	159
22 String Series Methods	161
22.1 The <code>value_counts</code> method	162
22.2 Special methods just for object columns	163
22.3 Exercises	166
23 Datetime Series Methods	169
23.1 The <code>dt</code> accessor	169
23.2 Datetime methods	170
23.3 Format time as a string with <code>strftime</code>	171
23.4 Convert to period	172
23.5 Timedeltas	173
23.6 Exercises	173
24 DataFrame Attributes and Methods	175
24.1 DataFrames vs Series	175
24.2 Core DataFrame Attributes	176
24.3 Arithmetic Operations with a DataFrame	178
24.4 Comparison Operators with DataFrames	179
24.5 Overlap of DataFrame and Series methods	179
24.6 Data Dictionaries	180
24.7 Exercises	181
25 DataFrame Statistical Methods	183
25.1 Differences between DataFrame and Series methods	183
25.2 Changing the direction of the operation	185
25.3 Non-Aggregation DataFrame methods	187
25.4 Nuisance Columns	191
25.5 Exercises	193
26 DataFrame Methods More	195
26.1 Methods for handling missing values	195
26.2 Filling missing values	196
26.3 Sorting	200
26.4 Finding the index of the maximum	203
26.5 Column and Row Dropping and Renaming	204
26.6 Adding a new column to the DataFrame	206
26.7 Methods Available only to Series and not DataFrames	207
26.8 Exercises	207
27 DataFrame Methods More II	209
27.1 Dropping duplicate rows	209
27.2 The <code>nunique</code> method	210
27.3 Addition and multiplication with string columns	211
27.4 The <code>clip</code> method	212
27.5 The <code>nlargest/nsmallest</code> methods	212
27.6 Copying a DataFrame	213
27.7 Inserting columns in the middle of a DataFrame	215
27.8 The <code>pop</code> method	217
27.9 The <code>corr</code> method	218
27.10 The <code>replace</code> method	218

27.11 Finding the maximum/minimum of a group	221
27.12 Exercises	223
28 Changing Data Types	225
28.1 Constructing a Series	225
28.2 numpy data types	225
28.3 Changing Data Types with astype	226
28.4 Float data types	227
28.5 Unsigned Integers	228
28.6 One size for booleans	228
28.7 Changing from float to int	228
28.8 Object data types	231
28.9 Setting data types in numpy arrays	232
28.10 The numpy datetime64 data type	233
28.11 The pandas datetime64 data type	234
28.12 Conversion failure	235
28.13 Force conversion with pd.to_numeric	235
28.14 DataFrame data type conversion	236
28.15 The astype method for DataFrames	238
28.16 Reading in data with known missing values	238
28.17 numpy and pandas timedelta64 data type	239
28.18 Period and Category data types	240
28.19 Different syntax for data types	240
28.20 Data Types Summary Table	240
28.21 Exercises	241
29 Assigning Subsets of Data	243
29.1 Setting new data with loc	243
29.2 Setting new data with iloc	245
29.3 Boolean Selection Assignment	246
29.4 Improper Assignment	247
29.5 Exercises	248
30 Case Study: Calculating Normality of Stock Market Returns	249
30.1 Registering for Quandl	249
30.2 Using the quandl package	249
30.3 Exercises	253
V Grouping Data	255
31 Groupby Aggregation Basics	257
31.1 Grouping with the groupby method	257
31.2 Syntax for using the groupby method	258
31.3 Grouping Column, Aggregating Column, Aggregating Function	259
31.4 More on method chaining with groupby	260
31.5 GroupBy objects	261
31.6 The index when grouping	262
31.7 Exercises	262
32 Grouping and Aggregating with Multiple Columns	265
32.1 Review grouping and aggregating with a single column	266
32.2 Grouping with Multiple Columns	266
32.3 Aggregating Multiple Columns	267
32.4 Grouping and Aggregating with Multiple Columns	268
32.5 Renaming all the columns	269
32.6 No added functionality of a MultiIndex	269
32.7 Multiple Grouping Columns, Aggregating Columns, and Aggregating Functions	269
32.8 Getting the size of each group	270
32.9 Rename the column when using reset_index	272

32.10 Exercises	272
33 Grouping with Pivot Tables	275
33.1 Creating the pivot table above with Pandas	275
33.2 Comparison to a group by	277
33.3 Styling pivot tables to find important data	279
33.4 Exercises	281
34 Counting with Crosstabs	283
34.1 Frequency counting with a Series	285
34.2 Counting the mental health occurrences by country	285
34.3 Which method to choose	287
34.4 Relative frequencies - only available with crosstab	287
34.5 Exercises	289
35 Alternate GroupBy Syntax	291
35.1 Grouping a single column, aggregating a single column, applying a single function	291
35.2 Multiple aggregation functions	292
35.3 Multiple Grouping, Aggregating, and Applying same Functions	293
35.4 A trick to discovering all the groupby methods	296
35.5 Exercises	297
36 Custom Aggregation	299
36.1 Using your customized aggregation function	299
36.2 Finding the percentage of all undergraduates represented in the top 5 most populous colleges	301
36.3 Optimizing a Custom Aggregation function	302
36.4 Pandas Power User Optimization	303
36.5 New Performance Test	306
36.6 Complexity vs Performance	306
36.7 Exercises	306
37 Transform and Filter with GroupBy	309
37.1 Back to boolean indexing	309
37.2 Shortcut with filter	311
37.3 Getting a nicer display	312
37.4 Using an anonymous function	312
37.5 The GroupBy transform method	314
37.6 transform must return either a single value or a Series the same length as the group	318
37.7 Summary of the GroupBy transform method	318
37.8 Exercises	318
38 Create Your Own Data Analysis	321
38.1 Overview	321
38.2 Begin Asking and Answering Questions	322
VI Time Series	323
39 Datetime and Timedelta	325
39.1 Date vs Time vs Datetime	325
39.2 Datetimes in NumPy	325
39.3 Pandas Timestamp	325
39.4 Why is to_datetime returning a Timestamp object?	326
39.5 Timestamp attributes	328
39.6 Timedelta - an amount of time	328
39.7 Timedelta attributes and methods	329
39.8 Creating Timedeltas by subtracting Datetimes	329
39.9 Exercises	331
40 Introduction to Time Series	333

40.1	Getting Stock Market Data	333
40.2	Set the Datetime column as the index	334
40.3	Easy subset selection with a DateTimeIndex	334
40.4	Slicing with partial string matching	335
40.5	Sampling Specific Times	335
40.6	Upsampling - Increasing the number of rows	337
40.7	Exercises	340
41	Grouping by Time	343
41.1	The PeriodIndex	345
41.2	The Period data type	345
41.3	Anchored offsets	347
41.4	Calling resample on a datetime column	348
41.5	Exercises	349
42	Rolling Windows	351
42.1	Visualization of Rolling Window	351
42.2	Keep window size the same with an integer	353
42.3	Plotting	355
42.4	Resampling and Rolling Windows with a Series - A bit easier	356
42.5	Exercises	358
43	Grouping by Time and another Column	359
43.1	Grouping by an amount of time and another column	361
43.2	Group independently	362
43.3	Using a pivot table with Grouper for easier comparisons	363
43.4	Exercises	365
44	Python datetime module	367
44.1	Create a date, a time and a datetime	367
44.2	Attributes of date, time, and datetimes	368
44.3	Methods of date, time, and datetimes	369
44.4	Alternate Constructors	369
44.5	Directives	370
44.6	Date and Datetime addition	371
44.7	Third-Party library dateutil	372
VII	Regular Expressions	375
45	Introduction to Regular Expressions	377
45.1	The contains and extract methods	377
45.2	Mini-Programming Language	378
45.3	Special Characters	379
45.4	The dot metacharacter	380
45.5	The caret metacharacter ^	380
45.6	The dollar sign metacharacter \$	381
45.7	Combining special characters	381
45.8	Exercises	382
46	Quantifiers	383
46.1	The asterisk metacharacter *	383
46.2	The plus metacharacter +	384
46.3	The question mark metacharacter ?	384
46.4	The curly braces metacharacter {m,n}	384
46.5	Exercises	385
47	Or Conditions	387
47.1	The pipe metacharacter	387
47.2	The brackets metacharacter []	388

47.3	Combining Special Characters	389
47.4	Exercises	390
48	Character Sets and Grouping	393
48.1	Special Characters lose their meaning within the brackets	393
48.2	The backslash \ metacharacter	394
48.3	The parentheses metacharacters ()	395
48.4	Using parentheses to change operator precedence	395
48.5	Using capture groups with the <code>extract</code> string method	397
48.6	Many other string methods take regexes	398
48.7	Other Dialects of Regex	399
48.8	More to Regex	399
48.9	Regex Summary	399
48.10	Exercises	400
49	Case Study - Explore Newsgroups with Regexes	401
49.1	Can you do the following?	402
50	Case Study - Feature Engineering on the Titanic	403
50.1	Exercises	403
VIII	Tidy Data	407
51	Tidy Data with <code>melt</code>	409
51.1	Tidy Data	409
51.2	Melting	410
51.3	Exercises	413
52	Reshaping by Pivoting	415
52.1	Inverting melted data with <code>pivot</code>	415
52.2	Pivoting with duplicate values	417
52.3	Using <code>pivot_table</code> to aggregate those values	420
52.4	Exercises	420
53	Common Messy Datasets	423
53.1	Most Common Messy Data Problems	423
53.2	Multiple variables are stored in one column	423
53.3	Two or more values are stored in the same cell	425
53.4	Variables are stored in both rows and columns	427
53.5	Steps to produce tidy data	431
53.6	Exercises	431
IX	Joining Data	433
54	Automatic Index Alignment	435
54.1	Adding two Series - Not as simple as it sounds	435
54.2	Adding a NumPy array to a Series	436
54.3	Adding Series that don't have the same index labels	437
54.4	Adding Series with duplicate values in the index	438
54.5	An exception to Cartesian Product rule	439
54.6	Cartesian product still happens if order is not the same	440
54.7	DataFrames align on both their index and columns	440
54.8	Cartesian Product over index and columns	443
55	Combining Data	445
55.1	Concatenating Data	445
55.2	Beware! Automatic Alignment of Index	448
55.3	Column names align first	448

55.4 Use axis=1 to change the direction of concatenation	448
56 SQL Databases	451
56.1 Connecting to a SQL database	451
56.2 The Chinook Database	451
56.3 Primary and Foreign Keys	452
56.4 Preparing the connection	452
56.5 Joining tables in Pandas with merge	454
56.6 Exercises	457
57 Data Normalization	459
57.1 Create a primary key to uniquely identify each row	460
57.2 We just created a dimension	461
57.3 Replacing original data with primary keys	463
57.4 Replace all the other dimensions with primary key columns	464
57.5 Fact Table	466
57.6 Data Model Diagram	466
57.7 Exercises	467
X Visualization	469
58 Matplotlib Fundamentals	471
58.1 Two Interfaces	471
58.2 Anatomy of a Figure	472
58.3 Embedding plots into Jupyter Notebooks	472
58.4 An aside on Unpacking	473
58.5 Back to Matplotlib	475
58.6 Verify the returned types of subplots	476
58.7 Setting the size of the Figure upon creation	477
58.8 Began the Object-Oriented Approach	478
58.9 Calling Axes methods - get_ and set_ methods	478
58.10 Use the tick_params method to set tick line properties	488
58.11 Exercise	490
59 Matplotlib Text and Lines	491
59.1 Add text to the Axes with the text method	491
59.2 Change the x and y limits to include the text	492
59.3 Set properties after creation	494
59.4 Adding horizontal and vertical lines	496
59.5 Create vertical lines with vlines	498
59.6 Add grid lines with the grid method	500
59.7 Annotating a point with an arrow	502
59.8 Exercise	503
60 Multiple Axes Figures	505
60.1 Create Multiple Axes with subplots	505
60.2 Distinguish the Figure from the Axes	506
60.3 Exercise	508
61 Matplotlib Data Plotting	509
61.1 The Axes API	509
61.2 The plot method - Creates line plots	509
61.3 Formatting the line	510
61.4 Matplotlib Colors	511
61.5 Integration with Pandas - plotting real data	512
61.6 Markers	514
61.7 Most common Plots	516
61.8 Univariate Analysis	516
61.9 Plotting dates	517

61.10	Plotting the Number of Riders per Day	519
61.11	Scatterplots	520
61.12	Creating a Legend	522
61.13	Exercises	522
62	Plotting with Pandas	525
62.1	Plotting a Series	525
62.2	Choosing other types of plots	528
62.3	Additional Plotting Arguments	529
62.4	Pandas plots return matplotlib objects	535
62.5	Exercises	537
63	Seaborn	539
63.1	Distribution Plots	541
63.2	Univariate Distribution Plots	541
63.3	Multivariate Distribution Plots	546
63.4	Adding an extra dimension with hue	550
63.5	Grouping and Aggregating Plots	552
63.6	Raw Plots	557
63.7	Heat maps	559
63.8	Grid Plots - Create Multiple Plots at once	561