

Master Data Analysis with Python

by
Ted Petrou

© 2019 Ted Petrou All Rights Reserved

Contents

I	Intro to pandas	1
1	What is pandas?	3
1.1	pandas operates on tabular data	4
1.2	pandas examples	4
1.3	Which pandas version to use?	5
1.4	Reading data	5
1.5	Filtering data	6
1.6	Aggregating methods	8
1.7	Non-aggregating methods	9
1.8	Aggregating within groups	9
1.9	Tidying	12
1.10	Joining Data	14
1.11	Time Series Analysis	15
1.12	Visualization	16
1.13	Much More	17
2	The DataFrame and Series	19
2.1	Reading external data with pandas	19
2.2	Components of a DataFrame	22
2.3	What type of object is <code>bikes</code> ?	23
2.4	Select a single column from a DataFrame - a Series	24
2.5	Components of a Series	24
2.6	Changing display options	25
2.7	Exercises	27
3	Data Types and Missing Values	29
3.1	Common data types	29
3.2	String data type - major enhancement to pandas 1.0	29
3.3	Missing value representation	30
3.4	New Integers and booleans data types in pandas 1.0	30
3.5	Recommendation for Pandas 1.0 - Avoid the new data types	30
3.6	Finding the data type of each column	31
3.7	Getting more metadata	32
3.8	More data types	34
3.9	Exercises	34
4	Setting a Meaningful Index	35
4.1	Setting an index of a DataFrame	35
4.2	Accessing the index, columns, and data	37
4.3	Accessing the components of a Series	39
4.4	Setting an index on read	40

4.5	Choosing a good index	42
4.6	Exercises	43
5	Five-Step Process for Data Exploration	45
II	Selecting Subsets of Data	49
6	Selecting Subsets of Data from DataFrames with just the brackets	51
6.1	pandas dual references: by label and by integer location	53
6.2	The three indexers [], loc, iloc	53
6.3	Begin with <i>just the brackets</i>	53
6.4	Select multiple columns with a list	54
6.5	Summary of <i>just the brackets</i>	56
6.6	Exercises	57
7	Selecting Subsets of Data from DataFrames with loc	59
7.1	Simultaneous row and column subset selection	59
7.2	loc with slice notation	61
7.3	Summary of the loc indexer	66
7.4	Exercises	66
8	Selecting Subsets of Data from DataFrames with iloc	69
8.1	Simultaneous row and column subset selection	69
8.2	Summary of iloc	74
8.3	Exercises	74
9	Selecting Subsets of Data from a Series	77
9.1	Series indexer rules	77
9.2	Series subset selection with loc	78
9.3	Series subset selection with iloc	79
9.4	Summary of Series subset selection	81
9.5	Exercises	81
10	Boolean Selection Single Conditions	85
10.1	Manually filtering the data	86
10.2	Operator overloading	87
10.3	Practical boolean selection	87
10.4	Boolean selection in one line	89
10.5	Single condition expression	89
10.6	Summary of single condition boolean selection	90
10.7	Exercises	90
11	Boolean Selection Multiple Conditions	93
11.1	Logical operators	93
11.2	Multiple conditions in one line	94
11.3	Using an or condition	95
11.4	Inverting a condition with the not operator	95
11.5	Many equality conditions in a single column	97
11.6	Exercises	99
12	Boolean Selection More	103
12.1	Boolean selection on a Series	103
12.2	The between method	105

12.3	Simultaneous boolean selection of rows and column labels with <code>loc</code>	105
12.4	Column to column comparisons	107
12.5	Finding Missing Values with <code>isna</code>	108
12.6	Exercises	108
13	Filtering with the query Method	111
13.1	The <code>query</code> method	111
13.2	Summary	118
13.3	Exercises	118
14	Miscellaneous Subset Selection	121
14.1	Selecting a column with dot notation	121
14.2	Slicing rows with just the brackets	123
14.3	Selecting a single cell with <code>at</code> and <code>iat</code>	124
III	Essential Series Commands	127
15	Series Attributes and Statistical Methods	129
15.1	Calling methods on a Series	129
15.2	Series Attributes and Methods	129
15.3	Core Series attributes	130
15.4	Arithmetic operators	131
15.5	Comparison operations	132
15.6	Boolean and bitwise operators	133
15.7	Statistical methods	133
15.8	Aggregation methods	134
15.9	Non-Aggregation methods	136
15.10	Series with a non-default index	138
15.11	Operations on a boolean Series	139
15.12	Exercises	140
16	Series Missing Value Methods	143
16.1	Methods for handling missing values	143
16.2	The <code>isna</code> method	143
16.3	Dropping missing values with <code>dropna</code>	145
16.4	Filling missing values with the <code>fillna</code> method	146
16.5	Filling missing values with <code>interpolate</code>	148
16.6	Graphing interpolation methods	149
16.7	Interpolation methods use the index	150
16.8	Exercises	151
17	Series Sorting, Ranking, and Uniqueness	155
17.1	Sorting	155
17.2	Ranking	157
17.3	Uniqueness	159
17.4	Exercises	161
18	Series Methods More	163
18.1	The <code>agg</code> method	163
18.2	Index of maximum and minimum	164
18.3	Differencing methods <code>diff</code> and <code>pct_change</code>	165
18.4	The <code>nlargest</code> and <code>nsmallest</code> methods	167

18.5 Randomly sample a Series	168
18.6 The <code>replace</code> method	169
18.7 Exercises	171
19 Series String Methods	173
19.1 The <code>value_counts</code> method	174
19.2 Special methods just for object columns	175
19.3 The <code>count</code> string method	176
19.4 The <code>contains</code> str method	176
19.5 The <code>len</code> str method	177
19.6 The <code>split</code> str method	177
19.7 The <code>replace</code> str method	178
19.8 Selecting substrings with the brackets	178
19.9 Regular expressions	179
19.10 Exercises	180
20 Series Datetime Methods	183
20.1 The <code>dt</code> accessor	183
20.2 Datetime Attributes	184
20.3 Datetime methods	186
20.4 Format time as a string with <code>strftime</code>	187
20.5 Convert to period	187
20.6 Timedeltas	189
20.7 Exercises	189
21 Project - Testing Normality of Stock Market Returns	191
21.1 Results discussion	194
21.2 Exercises	194
IV Essential DataFrame Commands	197
22 Introduction to DataFrame	199
22.1 DataFrames vs Series	199
22.2 Core DataFrame attributes	200
22.3 Arithmetic DataFrame operations	202
22.4 DataFrames comparison operators	205
22.5 Overlap of DataFrame and Series methods	205
22.6 Data Dictionaries	206
22.7 Exercises	207
23 DataFrame Statistical Methods	209
23.1 Aggregation methods	209
23.2 Changing the direction of the operation	211
23.3 Non-Aggregation methods	214
23.4 Nuisance Columns	217
23.5 Exercises	220
24 DataFrame Missing Value Methods	223
24.1 Methods for handling missing values	223
24.2 The <code>isna</code> method	223
24.3 Dropping rows and columns with the <code>dropna</code> method	225
24.4 Filling missing values with the <code>fillna</code> method	226

24.5	The <code>interpolate</code> method	229
24.6	Exercises	231
25	DataFrame Sorting, Ranking, and Uniqueness	233
25.1	Sorting	233
25.2	Ranking	236
25.3	Uniqueness	237
25.4	Finding the maximum/minimum of a group	239
25.5	Exercises	242
26	DataFrame Structure Methods	245
26.1	Adding a new column to the DataFrame	245
26.2	Copying a DataFrame	247
26.3	Column and Row Dropping and Renaming	250
26.4	Inserting columns in the middle of a DataFrame	253
26.5	The <code>pop</code> method	254
26.6	Exercises	255
27	DataFrame Methods More	257
27.1	The <code>agg</code> method	257
27.2	The <code>idxmax</code> and <code>idxmin</code> methods	258
27.3	The <code>diff</code> and <code>pct_change</code> methods	259
27.4	The <code>sample</code> method	260
27.5	The <code>nlargest</code> / <code>nsmallest</code> methods	261
27.6	The <code>corr</code> method	261
27.7	The <code>replace</code> method	262
27.8	Methods available only to Series and not DataFrames	266
27.9	Exercises	266
28	Assigning Subsets of Data	269
28.1	Setting new data with <code>loc</code>	269
28.2	Setting new data with <code>iloc</code>	271
28.3	Boolean selection assignment	272
28.4	Improper Assignment	273
28.5	Exercises	274
V	Data Types	275
29	Integer, Float, and Boolean Data types	277
29.1	Constructing a Series	277
29.2	Integer data type	277
29.3	Changing Data Types with <code>astype</code>	278
29.4	Unsigned Integers	279
29.5	Nullable integer data type	280
29.6	Summary of integer data type	283
29.7	Float data types	283
29.8	Changing from float to int	284
29.9	Boolean data type	286
29.10	Nullable boolean data type	288
29.11	Changing data types with an arithmetic operation	289
29.12	Setting data types in numpy arrays	290
29.13	Different syntax for data types	291

29.14	Boolean, integer, and float data type summary	292
29.15	Exercises	292
30	Object, String, and Categorical Data Types	295
30.1	Object data types	295
30.2	Categorical data type	297
30.3	Why the categorical data type is useful	299
30.4	The <code>cat</code> accessor	300
30.5	Modifying categories	300
30.6	Massive reduction in memory used	302
30.7	Speeding up operations	303
30.8	The <code>str</code> accessor is still available	304
30.9	Ordered categories	304
30.10	Integers can be categories	308
30.11	The new string data type	308
30.12	Converting strings to numeric	309
30.13	Force conversion with <code>pd.to_numeric</code>	310
30.14	Object, String, and Categorical data type summary	311
30.15	Exercises	311
30.16	2. Object, String, and Categorical Data Types	311
31	Datetime, Timedelta, and Period Data Types	315
31.1	The numpy <code>datetime64</code> data type	315
31.2	The pandas <code>datetime64</code> data type	317
31.3	The numpy <code>timedelta64</code> data type	319
31.4	The pandas <code>timedelta64</code> data type	319
31.5	The pandas <code>Period</code> data type	320
31.6	Datetime, Timedelta, and Period data type summary	321
31.7	Exercises	322
32	DataFrame Data Type Conversion	325
32.1	The <code>astype</code> method for DataFrames	327
32.2	Reading in data with known missing values	328
32.3	More data type conversion with the housing dataset	328
32.4	Exercises	330
VI	Grouping Data	333
33	Grouping Aggregation Basics	335
33.1	Group into independent DataFrames, then aggregate	335
33.2	Grouping with the <code>groupby</code> method	336
33.3	Syntax for using the <code>groupby</code> method	336
33.4	The index when grouping	338
33.5	More on method chaining with <code>groupby</code>	340
33.6	<code>GroupBy</code> objects	340
33.7	Exercises	341
34	Grouping and Aggregating with Multiple Columns	343
34.1	Review grouping and aggregating with a single column	343
34.2	Grouping with multiple columns	344
34.3	Aggregating multiple columns	345
34.4	Multiple grouping columns, aggregating columns, and aggregating functions	346

34.5	Getting the size of each group	346
34.6	Exercises	349
35	Grouping with Pivot Tables	351
35.1	Creating the pivot table above with pandas	351
35.2	Comparison to <code>groupby</code>	353
35.3	Styling pivot tables	354
35.4	Getting the size of each group	358
35.5	Add margins to get row and column totals	358
35.6	Non-standard pivot tables	359
35.7	Exercises	365
36	Counting with Crosstabs	367
36.1	Frequency counting with a Series	369
36.2	Counting the mental health occurrences by country	370
36.3	Counting frequency with the <code>crosstab</code> function	372
36.4	Exercises	375
37	Alternate Groupby Syntax	377
37.1	Aggregating a single column	377
37.2	No Aggregating Columns	380
38	Custom Aggregation	383
38.1	Using a customized aggregation function	383
38.2	Find the mean salary for the five highest paid employees per department	386
38.3	Percentage of employees by department with salaries greater than 100,000	389
38.4	Optimizing a Custom Aggregation function	391
38.5	Summary of Custom Aggregation Functions	396
38.6	Exercises	396
39	Transform and Filter with Groupby	399
39.1	The groupby <code>filter</code> method	399
39.2	Getting a nicer display	402
39.3	Finding actors that appear in at least 25 movies	404
39.4	Multiple conditions	405
39.5	The groupby <code>transform</code> method	406
39.6	Transforming multiple columns	410
39.7	Summary of the groupby <code>transform</code> method	412
39.8	Exercises	413
40	Other Groupby Methods	415
40.1	Kinds of groupby attributes and methods	415
40.2	Finding all available attributes and methods	417
40.3	Calling single aggregation methods	418
40.4	<code>head</code> , <code>tail</code> , and <code>nth</code> groupby methods	420
40.5	Non-aggregating methods	423
40.6	Exercises	426
41	Create Your Own Data Analysis	429
41.1	Overview	429
41.2	Begin Asking and Answering Questions	430

VII Time Series	431
42 Datetime and Timedelta	433
42.1 Date vs Time vs Datetime	433
42.2 Timestamp attributes	436
42.3 Timedelta - an amount of time	436
42.4 Timedelta attributes and methods	437
42.5 Creating Timedeltas by subtracting Datetimes	438
42.6 Exercises	439
43 Introduction to Time Series	441
43.1 Set the datetime column as the index	441
43.2 Easy subset selection with a DateTimeIndex	442
43.3 Sampling specific times	443
43.4 Upsampling - Increasing the number of rows	445
43.5 Exercises	448
44 Grouping by Time	451
44.1 The PeriodIndex	453
44.2 The Period data type	453
44.3 Anchored offsets	455
44.4 Calling <code>resample</code> on a datetime column	456
44.5 Calling <code>resample</code> on a Series	457
44.6 Exercises	458
45 Rolling Windows	461
45.1 Keep window size the same with an integer	463
45.2 Plotting	465
45.3 Exercises	466
46 Grouping by Time and another Column	467
46.1 Grouping by an amount of time and another column	469
46.2 Group independently	470
46.3 Using a pivot table with <code>Groupby</code> for easier comparisons	471
46.4 Exercises	473
VIII Regular Expressions	475
47 Introduction to Regular Expressions	477
47.1 The <code>contains</code> and <code>extract</code> methods	477
47.2 Mini-Programming Language	478
47.3 Special Characters	480
47.4 The dot metacharacter	480
47.5 The caret metacharacter <code>^</code>	480
47.6 The dollar sign metacharacter <code>\$</code>	481
47.7 Combining special characters	482
47.8 Exercises	482
48 Quantifiers	485
48.1 The asterisk metacharacter <code>*</code>	485
48.2 The plus metacharacter <code>+</code>	486
48.3 The question mark metacharacter <code>?</code>	486

48.4	The curly braces metacharacter <code>{m,n}</code>	487
48.5	Exercises	487
49	Or Conditions	489
49.1	The pipe metacharacter <code> </code>	489
49.2	The brackets metacharacter <code>[]</code>	490
49.3	Combining Special Characters	491
49.4	Exercises	492
50	Character Sets and Grouping	495
50.1	Special Characters lose their meaning within the brackets	495
50.2	The backslash <code>\</code> metacharacter	496
50.3	The parentheses metacharacters <code>()</code>	497
50.4	Using parentheses to change operator precedence	497
50.5	Using capture groups with the <code>extract</code> string method	499
50.6	Many other string methods take regexes	501
50.7	Other Dialects of Regex	501
50.8	More to Regex	501
50.9	Regex Summary	501
50.10	Exercises	502
51	Project - Explore Newsgroups with Regexes	505
51.1	Can you do the following?	506
52	Project - Feature Engineering on the Titanic	507
52.1	Exercises	507
IX	Tidy Data	511
53	Tidy Data with <code>melt</code>	513
53.1	Tidy Data	513
53.2	Melting	514
53.3	Exercises	517
54	Reshaping by Pivoting	519
54.1	Inverting melted data with <code>pivot</code>	519
54.2	Pivoting with duplicate values	521
54.3	Using <code>pivot_table</code> to aggregate those values	523
54.4	Exercises	523
55	Common Messy Datasets	525
55.1	Most common messy data problems	525
55.2	Multiple variables are stored in one column	525
55.3	Two or more values are stored in the same cell	527
55.4	Variables are stored in both rows and columns	529
55.5	Steps to produce tidy data	533
55.6	Exercises	533
X	Joining Data	535
56	Automatic Index Alignment	537
56.1	Adding two Series - Not as simple as it sounds	537

56.2	Adding a numpy array to a Series	538
56.3	Adding Series that don't have the same index labels	539
56.4	Adding Series with duplicate values in the index	540
56.5	An exception to Cartesian Product rule	542
56.6	DataFrames align on both their index and columns	543
56.7	Cartesian Product over index and columns	545
57	Combining Data	547
57.1	Concatenating Data	547
57.2	Beware! Automatic Alignment of Index	549
57.3	Column names align first	549
57.4	Use <code>axis=1</code> to change the direction of concatenation	550
58	SQL Databases	551
58.1	Connecting to a SQL database	551
58.2	The Chinook Database	551
58.3	Primary and Foreign Keys	552
58.4	Preparing the connection	552
58.5	Joining tables in Pandas with <code>merge</code>	554
58.6	Exercises	557
59	Data Normalization	559
59.1	Create a primary key to uniquely identify each row	560
59.2	We just created a dimension	561
59.3	Replacing original data with primary keys	563
59.4	Replace all the other dimensions with primary key columns	564
59.5	Fact Table	566
59.6	Data Model Diagram	566
59.7	Exercises	567
XI	Visualization	569
60	Matplotlib Fundamentals	571
60.1	Two Interfaces	571
60.2	Anatomy of a Figure	572
60.3	Embedding plots into Jupyter Notebooks	572
60.4	An aside on unpacking	573
60.5	Back to Matplotlib	576
60.6	Setting the size of the Figure upon creation	578
60.7	Began the Object-Oriented Approach	578
60.8	Calling Axes methods - <code>get_</code> and <code>set_</code> methods	579
60.9	Use the <code>tick_params</code> method to set tick line properties	588
60.10	Exercise	590
61	Matplotlib Text and Lines	591
61.1	Add text to the Axes with the <code>text</code> method	591
61.2	Change the x and y limits to include the text	592
61.3	Set properties after creation	593
61.4	Adding horizontal and vertical lines	595
61.5	Create vertical lines with <code>vlines</code>	596
61.6	Add grid lines with the <code>grid</code> method	599
61.7	Annotating a point with an arrow	600

62 Multiple Axes Figures	603
62.1 Create multiple axes with <code>subplots</code>	603
62.2 Distinguish the Figure from the Axes	604
63 Matplotlib Data Plotting	607
63.1 The Axes API	607
63.2 The <code>plot</code> method - Creates line plots	607
63.3 Formatting the line	608
63.4 Matplotlib Colors	609
63.5 Markers	611
63.6 Integration with Pandas - plotting real data	612
63.7 Most common Plots	613
63.8 Univariate Analysis	614
63.9 Plotting the Number of Riders per Day	616
63.10 Scatterplots	617
63.11 Creating a Legend	618
63.12 Exercises	619
64 Plotting with pandas	621
64.1 Plotting a Series	621
64.2 Choosing other types of plots	623
64.3 Additional plotting parameters	625
64.4 Pandas plots return matplotlib objects	630
64.5 Exercises	631
65 Seaborn	633
65.1 Distribution Plots	635
65.2 Univariate Distribution Plots	635
65.3 Multivariate Distribution Plots	639
65.4 Adding an extra dimension with <code>hue</code>	643
65.5 Grouping and Aggregating Plots	645
65.6 Raw Plots	650
65.7 Heat maps	652
65.8 Grid Plots - Create Multiple Plots at once	654