

Master Data Analysis with Python

by
Ted Petrou

© 2021 Ted Petrou All Rights Reserved

Contents

I	Intro to Pandas	17
1	What is pandas?	19
1.1	Why pandas and not xyz?	19
1.2	pandas operates on tabular data	20
1.3	pandas examples	20
1.4	Which pandas version to use?	21
1.5	Reading data	21
1.6	Filtering data	22
1.7	Aggregating methods	23
1.8	Non-aggregating methods	24
1.9	Aggregating within groups	24
1.10	Cleaning data	26
1.11	Joining Data	28
1.12	Time Series Analysis	29
1.13	Visualization	29
1.14	Much More	30
2	The DataFrame and Series	31
2.1	Reading external data with pandas	31
2.2	Components of a DataFrame	33
2.3	What type of object is bikes?	35
2.4	Select a single column from a DataFrame - a Series	36
2.5	Components of a Series	36
2.6	Changing display options	37
2.7	Exercises	38
3	Data Types and Missing Values	41
3.1	Common data types	41
3.2	String data type - major enhancement to pandas 1.0	41
3.3	Missing value representation	42
3.4	New Integers and booleans data types in pandas 1.0	42
3.5	Recommendation for Pandas 1.0 - Avoid the new data types	42
3.6	Finding the data type of each column	43
3.7	Getting more metadata	44
3.8	More data types	45
3.9	Exercises	45
4	Setting a Meaningful Index	47
4.1	Setting an index of a DataFrame	47
4.2	Accessing the index, columns, and data	49
4.3	Accessing the components of a Series	50

4.4	Setting an index on read	52
4.5	Choosing a good index	53
4.6	Exercises	54
5	Five-Step Process for Data Exploration	57
II Selecting Subsets of Data		61
6	Selecting Subsets of Data from DataFrames with just the brackets	63
6.1	pandas dual references - by label and by integer location	65
6.2	The three indexers [], loc, iloc	65
6.3	Begin with <i>just the brackets</i>	66
6.4	Select multiple columns with a list	66
6.5	Summary of <i>just the brackets</i>	68
6.6	Exercises	68
7	Selecting Subsets of Data from DataFrames with loc	71
7.1	Simultaneous row and column subset selection	71
7.2	loc with slice notation	73
7.3	Summary of the loc indexer	77
7.4	Exercises	77
8	Selecting Subsets of Data from DataFrames with iloc	81
8.1	Simultaneous row and column subset selection	81
8.2	Summary of iloc	85
8.3	Exercises	85
9	Selecting Subsets of Data from a Series	87
9.1	Series indexer rules	87
9.2	Use loc and iloc instead of just the brackets	88
9.3	Series subset selection with loc	88
9.4	Series subset selection with iloc	89
9.5	Summary of Series subset selection	91
9.6	Exercises	91
10	Boolean Selection Single Conditions	93
10.1	Manually filtering the data	94
10.2	Operator overloading	95
10.3	Practical boolean selection	95
10.4	Boolean selection in one line	97
10.5	Single condition expression	97
10.6	Summary of single condition boolean selection	97
10.7	Exercises	97
11	Boolean Selection Multiple Conditions	101
11.1	Logical operators	101
11.2	Multiple conditions in one line	102
11.3	Using an or condition	103
11.4	Inverting a condition with the not operator	103
11.5	Many equality conditions in a single column	104
11.6	Exercises	105

12 Boolean Selection More	107
12.1 Boolean selection on a Series	107
12.2 The <code>between</code> method	109
12.3 Simultaneous boolean selection of rows and column labels with <code>loc</code>	109
12.4 Column to column comparisons	110
12.5 Finding missing values with <code>isna</code>	111
12.6 Exercises	111
13 Filtering with the query Method	115
13.1 The <code>query</code> method	115
13.2 Use strings <code>and</code> , <code>or</code> , <code>not</code>	116
13.3 Chained comparisons	116
13.4 Reference strings with quotes	116
13.5 Column to column comparisons	117
13.6 Use 'in' for multiple equalities	117
13.7 Arithmetic operations within <code>query</code>	117
13.8 Reference variable names with the <code>@</code> symbol	118
13.9 Using the index with <code>query</code>	119
13.10 Use backticks to reference column names with spaces	119
13.11 Summary	120
13.12 Exercises	120
14 Miscellaneous Subset Selection	123
14.1 Selecting a column with dot notation	123
14.2 Selecting rows with just the brackets using slice notation	125
14.3 Selecting a single cell with <code>at</code> and <code>iat</code>	126
14.4 Exercises	127
III Essential Series Commands	129
15 Numeric Series Methods	131
15.1 Calling Series methods	131
15.2 City of Houston Employee Data	132
15.3 Core Series attributes	132
15.4 Arithmetic operators	133
15.5 Comparison operations	134
15.6 Boolean and bitwise operators	135
15.7 Statistical methods	135
15.8 Aggregation methods	136
15.9 Non-Aggregation methods	138
15.10 Series methods with a non-default index	140
15.11 Operations on a boolean Series	141
15.12 Exercises	142
16 Series Missing Value Methods	145
16.1 Methods for handling missing values	145
16.2 The <code>isna</code> method	145
16.3 Dropping missing values with <code>dropna</code>	147
16.4 Filling missing values with the <code>fillna</code> method	148
16.5 Filling missing values with <code>interpolate</code>	150
16.6 Graphing interpolation methods	151
16.7 Interpolation methods use the index	152

16.8 Exercises	152
17 Series Sorting, Ranking, and Uniqueness	155
17.1 Sorting	155
17.2 Ranking	157
17.3 Uniqueness	159
17.4 Exercises	161
18 More Series Methods	163
18.1 The <code>agg</code> method	163
18.2 The index of the minimum and maximum	164
18.3 Differencing methods <code>diff</code> and <code>pct_change</code>	165
18.4 The <code>nsmallest</code> and <code>nlargest</code> methods	167
18.5 Randomly sample a Series	168
18.6 The <code>replace</code> method	169
18.7 Exercises	171
19 String Series 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	179
20 Datetime Series 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	188
20.7 Exercises	189
21 Project - Testing Normality of Stock Market Returns	191
21.1 Results discussion	193
21.2 Exercises	194
IV Essential DataFrame Commands	197
22 Introduction to DataFrames	199
22.1 DataFrames vs Series	199
22.2 Core DataFrame attributes	200
22.3 Arithmetic DataFrame operations	202
22.4 DataFrame comparison operators	204
22.5 Overlap of DataFrame and Series methods	204
22.6 Data Dictionaries	205
22.7 Exercises	206

23 Numeric DataFrame Methods	207
23.1 Aggregation methods	207
23.2 Changing the direction of the operation	209
23.3 Non-Aggregation methods	211
23.4 Nuisance Columns	215
23.5 Exercises	217
24 DataFrame Missing Value Methods	221
24.1 Methods for handling missing values	221
24.2 The <code>isna</code> method	221
24.3 Dropping rows and columns with the <code>dropna</code> method	223
24.4 Filling missing values with the <code>fillna</code> method	224
24.5 The <code>interpolate</code> method	227
24.6 Exercises	228
25 DataFrame Sorting, Ranking, and Uniqueness	231
25.1 Sorting	231
25.2 Ranking	233
25.3 Uniqueness	234
25.4 Finding the maximum/minimum of a group	236
25.5 Exercises	238
26 DataFrame Structure Methods	241
26.1 Adding a new column to the DataFrame	241
26.2 Copying a DataFrame	243
26.3 Column and Row Dropping and Renaming	245
26.4 Inserting columns in the middle of a DataFrame	247
26.5 The <code>pop</code> method	248
26.6 Exercises	249
27 More DataFrame Methods	251
27.1 The <code>agg</code> method	251
27.2 The index of the minimum and maximum	252
27.3 Differencing methods <code>diff</code> and <code>pct_change</code>	252
27.4 The <code>sample</code> method	253
27.5 The <code>nsmallest</code> and <code>nlargest</code> methods	254
27.6 The <code>corr</code> method	255
27.7 The <code>replace</code> method	256
27.8 Methods available only to Series and not DataFrames	258
27.9 Exercises	258
28 Assigning Subsets of Data	259
28.1 Setting new data with <code>loc</code>	259
28.2 Setting new data with <code>iloc</code>	261
28.3 Boolean selection assignment	262
28.4 Improper Assignment	262
28.5 Exercises	264
V Data Types	265
29 Integer, Float, and Boolean Data types	267
29.1 Constructing a Series	267

29.2 Integer data type	267
29.3 Changing data types with <code>astype</code>	268
29.4 Unsigned Integers	269
29.5 Nullable integer data type	270
29.6 Float data types	273
29.7 Changing from float to int	274
29.8 pandas nullable float data type	275
29.9 Boolean data type	275
29.10 Nullable boolean data type	277
29.11 Changing data types with an arithmetic operation	278
29.12 Setting data types in numpy arrays	280
29.13 Different syntax for data types	280
29.14 Boolean, integer, and float data type summary	282
29.15 Exercises	282
30 Object, Categorical, and String Data Types	285
30.1 Object data types	285
30.2 Categorical data type	287
30.3 Why the categorical data type is useful	289
30.4 The <code>cat</code> accessor	290
30.5 Modifying categories	290
30.6 Massive reduction in memory used	292
30.7 Speeding up operations	293
30.8 The <code>str</code> accessor is still available	294
30.9 Ordered categories	294
30.10 Integers can be categories	298
30.11 The new string data type	298
30.12 Converting strings to numeric	299
30.13 Force conversion with <code>pd.to_numeric</code>	300
30.14 Object, String, and Categorical data type summary	301
30.15 Exercises	301
31 Datetime, Timedelta, and Period Data Types	305
31.1 Definitions	305
31.2 The numpy <code>datetime64</code> data type	305
31.3 The pandas <code>datetime64</code> data type	307
31.4 The numpy <code>timedelta64</code> data type	309
31.5 The pandas <code>timedelta64</code> data type	309
31.6 The pandas period data type	310
31.7 Datetime, Timedelta, and Period data type summary	311
31.8 Exercises	312
32 DataFrame Data Type Conversion	315
32.1 The <code>astype</code> method for DataFrames	317
32.2 Reading in data with known missing values	317
32.3 More data type conversion with the housing dataset	318
32.4 Exercises	320
VI Grouping Data	321
33 Grouping Aggregation Basics	323
33.1 Group into independent DataFrames, then aggregate	323

33.2	Grouping with the <code>groupby</code> method	324
33.3	Syntax for using the <code>groupby</code> method	324
33.4	The index when grouping	326
33.5	More on method chaining with <code>groupby</code>	327
33.6	<code>GroupBy</code> objects	328
33.7	Exercises	329
34	Grouping and Aggregating with Multiple Columns	331
34.1	Review grouping and aggregating with a single column	331
34.2	Grouping with multiple columns	332
34.3	Aggregating multiple columns	333
34.4	Multiple grouping columns, aggregating columns, and aggregating functions	333
34.5	Getting the size of each group	334
34.6	Exercises	336
35	Grouping with Pivot Tables	339
35.1	Creating pivot tables with pandas	339
35.2	Comparison to <code>groupby</code>	341
35.3	Styling pivot tables	342
35.4	Getting the size of each group	345
35.5	Add margins to get row and column totals	346
35.6	Non-standard pivot tables	346
35.7	Exercises	350
36	Counting with Crosstabs	353
36.1	Frequency counting with a Series	355
36.2	Counting the mental health occurrences by country	356
36.3	Counting frequency with the <code>crosstab</code> function	357
36.4	Exercises	360
37	Alternative Groupby Syntax	361
37.1	Aggregating a single column	361
37.2	No Aggregating Columns	363
37.3	Exercises	364
38	Custom Aggregation	367
38.1	Using a customized aggregation function	367
38.2	Find the mean salary for the five highest paid employees per department	370
38.3	Percentage of employees by department with salaries greater than 100,000	372
38.4	Optimizing a Custom Aggregation function	374
38.5	Summary of Custom Aggregation Functions	378
38.6	Exercises	378
39	Transform and Filter with Groupby	381
39.1	The <code>groupby filter</code> method	381
39.2	Getting a nicer display	384
39.3	Finding actors that appear in at least 25 movies	385
39.4	Multiple conditions	386
39.5	The <code>groupby transform</code> method	387
39.6	Transforming multiple columns	391
39.7	Summary of the <code>groupby transform</code> method	392
39.8	Exercises	393

40 More Groupby Methods	395
40.1 Kinds of groupby attributes and methods	395
40.2 Finding all available attributes and methods	397
40.3 Calling single aggregation methods	398
40.4 <code>head</code> , <code>tail</code> , and <code>nth</code> groupby methods	399
40.5 Non-aggregating methods	402
40.6 Exercises	404
41 Binning Numeric Columns	407
41.1 Grouping with numeric columns	407
41.2 Binning with <code>pd.cut</code>	408
41.3 Quantile binning with <code>pd.qcut</code>	411
41.4 Grouping with bins	411
42 Create Your Own Data Analysis	413
42.1 Overview	413
42.2 Begin Asking and Answering Questions	414
VII Time Series	415
43 Datetime, Timedelta, and Period Objects	417
43.1 Definitions	417
43.2 Date vs Time vs Datetime	417
43.3 Creating single datetime objects in pandas	418
43.4 Timestamp attributes and methods	422
43.5 Creating single timedelta objects in pandas	424
43.6 Timedelta attributes and methods	425
43.7 Creating timedeltas by subtracting datetimes	426
43.8 Creating Period Objects in Pandas	427
43.9 Creating multiple datetimes and timestamps	428
43.10 Exercises	428
44 Selecting Time Series Data	431
44.1 Set the datetime column as the index	431
44.2 Easy subset selection with a <code>DateTimeIndex</code>	432
44.3 Selecting rows at specific frequencies	433
44.4 Upsampling - Increasing the number of rows	435
44.5 Use integers in the offset alias	436
44.6 Exercises	438
45 Grouping by Time	441
45.1 Grouping with the <code>resample</code> method	441
45.2 Grouping by different time periods	443
45.3 Grouping by more than one consecutive offset alias period	445
45.4 Grouping by time with the <code>groupby</code> method	446
45.5 Calling <code>resample</code> on a datetime column	448
45.6 Calling <code>resample</code> on a Series	448
45.7 Exercises	449
46 Rolling Windows	453
46.1 The <code>rolling</code> method	453
46.2 Rolling with offset aliases	455

46.3 DataFrame rolling method	457
46.4 Exercises	458
47 Grouping by Time and another Column	461
47.1 Grouping by an amount of time and another column	462
47.2 Group independently	462
47.3 Using a pivot table with Grouper for easier comparisons	463
47.4 Rolling windows within a group	463
47.5 Exercises	465
48 More Time Series Functionality	467
48.1 Selecting multiple rows at specific frequencies	467
48.2 Shifting the data	471
48.3 Creating date ranges	475
48.4 Exercises	476
VIII Regular Expressions	479
49 Introduction to Regular Expressions	481
49.1 The contains and extract methods	481
49.2 Mini-Programming Language	482
49.3 Special Characters	484
49.4 The dot metacharacter	484
49.5 The caret metacharacter <code>^</code>	484
49.6 The dollar sign metacharacter <code>\$</code>	485
49.7 Combining special characters	486
49.8 Exercises	486
50 Quantifiers	489
50.1 The asterisk metacharacter <code>*</code>	489
50.2 The plus metacharacter <code>+</code>	490
50.3 The question mark metacharacter <code>?</code>	490
50.4 The curly braces metacharacter <code>{m,n}</code>	491
50.5 Exercises	491
51 Or Conditions	493
51.1 The pipe metacharacter <code> </code>	493
51.2 The brackets metacharacter <code>[]</code>	494
51.3 Combining Special Characters	495
51.4 Exercises	496
52 Character Sets and Grouping	499
52.1 Special Characters lose their meaning within the brackets	499
52.2 The backslash <code>\</code> metacharacter	500
52.3 The parentheses metacharacters <code>()</code>	501
52.4 Using parentheses to change operator precedence	501
52.5 Using capture groups with the extract string method	503
52.6 Many other string methods take regexes	505
52.7 Other Dialects of Regex	505
52.8 More to Regex	505
52.9 Regex Summary	505
52.10 Exercises	506

53 Project - Explore Newsgroups with Regexes	509
53.1 Can you do the following?	510
54 Project - Feature Engineering on the Titanic	511
54.1 Exercises	511
IX Tidy Data	515
55 Tidy Data with melt	517
55.1 Tidy data and Lego	517
55.2 Tidy Data	517
55.3 Melting	518
55.4 Exercises	520
56 Reshaping by Pivoting	523
56.1 Inverting melted data with <code>pivot</code>	523
56.2 Pivoting with duplicate values	524
56.3 Using <code>pivot_table</code> to aggregate those values	526
56.4 Exercises	526
57 Common Messy Datasets	529
57.1 Most common messy data problems	529
57.2 Multiple variables are stored in one column	529
57.3 Two or more values are stored in the same cell	531
57.4 Variables are stored in both rows and columns	533
57.5 Steps to produce tidy data	536
57.6 Exercises	537
X Joining Data	539
58 Automatic Index Alignment	541
58.1 Adding two Series - Not as simple as it sounds	541
58.2 Adding together numpy arrays	543
58.3 Operating on two Series with different index values	544
58.4 Adding Series with duplicate values in the index	545
58.5 An exception to Cartesian Product rule	546
58.6 Arithmetic operations with two DataFrames	547
58.7 Adding new columns to DataFrame from a Series	549
58.8 Arithmetic operations with one DataFrame and one Series	552
58.9 Exercises	555
59 Concatenating Data	557
59.1 Concatenation with the <code>pd.concat</code> function	557
59.2 Automatic alignment of index always takes place	560
59.3 Change the direction of concatenation with <code>axis</code>	560
59.4 Appending rows to DataFrames	561
59.5 Adding new columns with <code>assign</code>	562
60 SQL Databases	565
60.1 Connecting to a SQL database	565
60.2 The Chinook Database	565
60.3 Primary and Foreign Keys	566

60.4	Preparing the connection	566
60.5	Joining tables in Pandas with <code>merge</code>	567
60.6	Exercises	569
61	Data Normalization	571
61.1	Create a primary key to uniquely identify each row	572
61.2	We just created a dimension	573
61.3	Replacing original data with primary keys	575
61.4	Replace all the other dimensions with primary key columns	575
61.5	Fact Table	577
61.6	Data Model Diagram	577
61.7	Exercises	578
XI	Visualization with Matplotlib	579
62	Introduction to matplotlib	581
62.1	Two interfaces of matplotlib	581
62.2	Figure - Axes Hierarchy	582
62.3	Setting the size of the figure upon creation	585
62.4	Axes methods	586
62.5	Change tick label and tick line properties with <code>tick_params</code>	593
62.6	Setting multiple properties at the same time with <code>set</code>	594
62.7	Exercises	595
63	Matplotlib Text and Lines	597
63.1	The axes <code>text</code> method	597
63.2	Creating horizontal lines with <code>hlines</code>	603
63.3	Create vertical lines with <code>vlines</code>	605
63.4	Add grid lines with the <code>grid</code> method	607
63.5	Aligning text horizontally and vertically	609
63.6	Add text with arrows using the <code>annotate</code> method	611
63.7	Exercises	615
64	Matplotlib Resolution	617
64.1	Matplotlib inches	617
64.2	Creating figures with custom DPI	620
64.3	Text and line “points”	621
64.4	Run configuration settings	622
64.5	Creating style sheets	626
64.6	Exercises	629
65	Matplotlib Patches and Colors	631
65.1	Adding matplotlib patches	631
65.2	Circle patches	631
65.3	Ellipse patches	633
65.4	Rectangle patches	634
65.5	Polygon patches	634
65.6	Arc patches	635
65.7	Wedge patches	636
65.8	Matplotlib colors	638
65.9	Color transparency	643
65.10	Layering with <code>zorder</code>	644

65.11	Gray scale	645
65.12	Colormaps	646
65.13	Filling between two lines	648
65.14	Creating a basketball court	650
65.15	Exercises	654
66	Matplotlib Line Plots	657
66.1	Axes API	657
66.2	Line plots with the <code>plot</code> method	657
66.3	Integration with pandas	660
66.4	Color cycle	664
66.5	More line plots	665
66.6	Adding a legend	668
66.7	Exercises	673
67	Matplotlib Scatter and Bar Plots	675
67.1	Scatter plots	675
67.2	Change scatter plot point size	685
67.3	Bar plots	688
67.4	Exercises	696
68	Matplotlib Distribution Plots	697
68.1	Histograms	698
68.2	Box and whisker plots	704
68.3	Exercises	706
69	Best of the Rest of Matplotlib	707
69.1	Axes spines	707
69.2	The <code>xaxis</code> and <code>yaxis</code> objects	709
69.3	Tick locators	710
69.4	Tick formatters	711
69.5	Minor ticks	712
69.6	Horizontal and vertical lines that span the axes	713
69.7	Plotting with dates	714
69.8	Using a different scale for the axis	717
69.9	Adding images	721
69.10	Coordinate systems	726
69.11	Figure methods	729
69.12	Creating a grid of axes	732
69.13	Exercises	735
XII	Visualization with Pandas and Seaborn	737
70	Plotting with pandas Series	739
70.1	Line plots	740
70.2	Bar plots	742
70.3	Distribution plots	743
70.4	Pie Charts	744
70.5	Area Plots	745
70.6	Adding a plot to a previously made axes	745
71	Plotting with pandas DataFrames	747

71.1	Line plots	747
71.2	Bar plots	749
71.3	Plotting on separate axes	752
71.4	Distribution plots	753
71.5	Scatter and Hexbin	754
71.6	Area plots	758
72	Seaborn Axes Plots	765
72.1	The seaborn API	765
72.2	seaborn integration with pandas	766
72.3	Distribution Plots	766
72.4	Seaborn style sheets	771
72.5	Other distribution plots	771
72.6	Automatic grouping by category	772
72.7	Grouping within groups with <code>hue</code>	774
72.8	Tidy data	775
72.9	Grouping and Aggregating Plots	776
72.10	Raw data plots	784
72.11	Scatter plots with linear regression lines using <code>regplot</code>	789
72.12	Ordered categorical data	791
72.13	Exercises	793
73	Seaborn Grid Plots	795
73.1	Grids by categories	795
73.2	Scatter and line plot grids	799
73.3	Regression grid plots	801
73.4	Bivariate distributions grids	802
73.5	Scatter plot grids of multiple column combinations	803
73.6	Hierarchichal cluster map	804