

# Master Data Analysis with Python

by  
Ted Petrou

© 2021 Ted Petrou All Rights Reserved



# Contents

<b>I</b>	<b>Intro to Pandas</b>	<b>17</b>
<b>1</b>	<b>What is pandas?</b>	<b>19</b>
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
<b>2</b>	<b>The DataFrame and Series</b>	<b>31</b>
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
<b>3</b>	<b>Data Types and Missing Values</b>	<b>41</b>
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
<b>4</b>	<b>Setting a Meaningful Index</b>	<b>47</b>
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
<b>5</b>	<b>Five-Step Process for Data Exploration</b>	<b>57</b>
<b>II Selecting Subsets of Data</b>		<b>61</b>
<b>6</b>	<b>Selecting Subsets of Data from DataFrames with just the brackets</b>	<b>63</b>
6.1	pandas dual references - by label and by integer location . . . . .	64
6.2	The three indexers [ ], loc, iloc . . . . .	65
6.3	Begin with <i>just the brackets</i> . . . . .	65
6.4	Select multiple columns with a list . . . . .	66
6.5	Summary of <i>just the brackets</i> . . . . .	68
6.6	Exercises . . . . .	68
<b>7</b>	<b>Selecting Subsets of Data from DataFrames with loc</b>	<b>69</b>
7.1	Simultaneous row and column subset selection . . . . .	69
7.2	loc with slice notation . . . . .	71
7.3	Summary of the loc indexer . . . . .	75
7.4	Exercises . . . . .	75
<b>8</b>	<b>Selecting Subsets of Data from DataFrames with iloc</b>	<b>79</b>
8.1	Simultaneous row and column subset selection . . . . .	79
8.2	Summary of iloc . . . . .	83
8.3	Exercises . . . . .	83
<b>9</b>	<b>Selecting Subsets of Data from a Series</b>	<b>85</b>
9.1	Series indexer rules . . . . .	85
9.2	Use loc and iloc instead of just the brackets . . . . .	86
9.3	Series subset selection with loc . . . . .	86
9.4	Series subset selection with iloc . . . . .	87
9.5	Summary of Series subset selection . . . . .	89
9.6	Exercises . . . . .	89
<b>10</b>	<b>Boolean Selection Single Conditions</b>	<b>91</b>
10.1	Manually filtering the data . . . . .	92
10.2	Operator overloading . . . . .	93
10.3	Practical boolean selection . . . . .	93
10.4	Boolean selection in one line . . . . .	95
10.5	Single condition expression . . . . .	95
10.6	Summary of single condition boolean selection . . . . .	95
10.7	Exercises . . . . .	95
<b>11</b>	<b>Boolean Selection Multiple Conditions</b>	<b>99</b>
11.1	Logical operators . . . . .	99
11.2	Multiple conditions in one line . . . . .	100
11.3	Using an or condition . . . . .	101
11.4	Inverting a condition with the not operator . . . . .	101
11.5	Many equality conditions in a single column . . . . .	102
11.6	Exercises . . . . .	103

<b>12 Boolean Selection More</b>	<b>105</b>
12.1 Boolean selection on a Series	105
12.2 The <code>between</code> method	107
12.3 Simultaneous boolean selection of rows and column labels with <code>loc</code>	107
12.4 Column to column comparisons	108
12.5 Finding missing values with <code>isna</code>	109
12.6 Exercises	109
<b>13 Filtering with the query Method</b>	<b>113</b>
13.1 The <code>query</code> method	113
13.2 Use strings <code>and</code> , <code>or</code> , <code>not</code>	114
13.3 Chained comparisons	114
13.4 Reference strings with quotes	114
13.5 Column to column comparisons	115
13.6 Use 'in' for multiple equalities	115
13.7 Arithmetic operations within <code>query</code>	115
13.8 Reference variable names with the <code>@</code> symbol	116
13.9 Using the index with <code>query</code>	117
13.10 Use backticks to reference column names with spaces	117
13.11 Summary	118
13.12 Exercises	118
<b>14 Miscellaneous Subset Selection</b>	<b>121</b>
14.1 Selecting a column with dot notation	121
14.2 Selecting rows with just the brackets using slice notation	123
14.3 Selecting a single cell with <code>at</code> and <code>iat</code>	124
14.4 Exercises	125
<b>III Essential Series Commands</b>	<b>127</b>
<b>15 Numeric Series Methods</b>	<b>129</b>
15.1 Calling Series methods	129
15.2 City of Houston Employee Data	130
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 methods with a non-default index	138
15.11 Operations on a boolean Series	139
15.12 Exercises	140
<b>16 Series Missing Value Methods</b>	<b>143</b>
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	150
<b>17 Series Sorting, Ranking, and Uniqueness</b>	<b>153</b>
17.1 Sorting	153
17.2 Ranking	155
17.3 Uniqueness	157
17.4 Exercises	159
<b>18 More Series Methods</b>	<b>161</b>
18.1 The <code>agg</code> method	161
18.2 The index of the minimum and maximum	162
18.3 Differencing methods <code>diff</code> and <code>pct_change</code>	163
18.4 The <code>nsmallest</code> and <code>nlargest</code> methods	165
18.5 Randomly sample a Series	166
18.6 The <code>replace</code> method	167
18.7 Exercises	169
<b>19 String Series Methods</b>	<b>171</b>
19.1 The <code>value_counts</code> method	172
19.2 Special methods just for object columns	173
19.3 The <code>count</code> string method	174
19.4 The <code>contains</code> str method	174
19.5 The <code>len</code> str method	175
19.6 The <code>split</code> str method	175
19.7 The <code>replace</code> str method	176
19.8 Selecting substrings with the brackets	176
19.9 Regular expressions	177
19.10 Exercises	177
<b>20 Datetime Series Methods</b>	<b>181</b>
20.1 The <code>dt</code> accessor	181
20.2 Datetime Attributes	182
20.3 Datetime methods	184
20.4 Format time as a string with <code>strftime</code>	185
20.5 Convert to period	185
20.6 Timedeltas	187
20.7 Exercises	187
<b>21 Project - Testing Normality of Stock Market Returns</b>	<b>189</b>
21.1 Results discussion	191
21.2 Exercises	192
<b>IV Essential DataFrame Commands</b>	<b>195</b>
<b>22 Introduction to DataFrames</b>	<b>197</b>
22.1 DataFrames vs Series	197
22.2 Core DataFrame attributes	198
22.3 Arithmetic DataFrame operations	200
22.4 DataFrame comparison operators	202
22.5 Overlap of DataFrame and Series methods	202
22.6 Data Dictionaries	203
22.7 Exercises	204

<b>23 Numeric DataFrame Methods</b>	<b>205</b>
23.1 Aggregation methods . . . . .	205
23.2 Changing the direction of the operation . . . . .	207
23.3 Non-Aggregation methods . . . . .	209
23.4 Nuisance Columns . . . . .	213
23.5 Exercises . . . . .	215
<b>24 DataFrame Missing Value Methods</b>	<b>219</b>
24.1 Methods for handling missing values . . . . .	219
24.2 The <code>isna</code> method . . . . .	219
24.3 Dropping rows and columns with the <code>dropna</code> method . . . . .	221
24.4 Filling missing values with the <code>fillna</code> method . . . . .	222
24.5 The <code>interpolate</code> method . . . . .	225
24.6 Exercises . . . . .	226
<b>25 DataFrame Sorting, Ranking, and Uniqueness</b>	<b>229</b>
25.1 Sorting . . . . .	229
25.2 Ranking . . . . .	231
25.3 Uniqueness . . . . .	232
25.4 Finding the maximum/minimum of a group . . . . .	234
25.5 Exercises . . . . .	236
<b>26 DataFrame Structure Methods</b>	<b>239</b>
26.1 Adding a new column to the DataFrame . . . . .	239
26.2 Copying a DataFrame . . . . .	241
26.3 Column and Row Dropping and Renaming . . . . .	243
26.4 Inserting columns in the middle of a DataFrame . . . . .	245
26.5 The <code>pop</code> method . . . . .	246
26.6 Exercises . . . . .	247
<b>27 More DataFrame Methods</b>	<b>249</b>
27.1 The <code>agg</code> method . . . . .	249
27.2 The index of the minimum and maximum . . . . .	250
27.3 Differencing methods <code>diff</code> and <code>pct_change</code> . . . . .	250
27.4 The <code>sample</code> method . . . . .	251
27.5 The <code>nsmallest</code> and <code>nlargest</code> methods . . . . .	252
27.6 The <code>corr</code> method . . . . .	253
27.7 The <code>replace</code> method . . . . .	254
27.8 Methods available only to Series and not DataFrames . . . . .	256
27.9 Exercises . . . . .	256
<b>28 Assigning Subsets of Data</b>	<b>259</b>
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
<b>V Data Types</b>	<b>265</b>
<b>29 Integer, Float, and Boolean Data types</b>	<b>267</b>
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
<b>30 Object, Categorical, and String Data Types</b>	<b>285</b>
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
<b>31 Datetime, Timedelta, and Period Data Types</b>	<b>305</b>
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 . . . . .	312
31.8 Exercises . . . . .	312
<b>32 DataFrame Data Type Conversion</b>	<b>315</b>
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
<b>VI Grouping Data</b>	<b>321</b>
<b>33 Grouping Aggregation Basics</b>	<b>323</b>
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
<b>34</b>	<b>Grouping and Aggregating with Multiple Columns</b>	<b>331</b>
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
<b>35</b>	<b>Grouping with Pivot Tables</b>	<b>339</b>
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
<b>36</b>	<b>Counting with Crosstabs</b>	<b>353</b>
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	Normalizing other aggregations	360
36.5	Exercises	361
<b>37</b>	<b>Alternative Groupby Syntax</b>	<b>363</b>
37.1	Aggregating a single column	363
37.2	No Aggregating Columns	365
37.3	Exercises	366
<b>38</b>	<b>Custom Aggregation</b>	<b>369</b>
38.1	Using a customized aggregation function	369
38.2	Find the mean salary for the five highest paid employees per department	372
38.3	Percentage of employees by department with salaries greater than 100,000	374
38.4	Optimizing a Custom Aggregation function	376
38.5	Summary of Custom Aggregation Functions	380
38.6	Exercises	380
<b>39</b>	<b>Transform and Filter with Groupby</b>	<b>383</b>
39.1	The <code>groupby filter</code> method	383
39.2	Getting a nicer display	386
39.3	Finding actors that appear in at least 25 movies	387
39.4	Multiple conditions	388
39.5	The <code>groupby transform</code> method	389
39.6	Transforming multiple columns	393
39.7	Summary of the <code>groupby transform</code> method	394
39.8	Exercises	395

<b>40 More Groupby Methods</b>	<b>397</b>
40.1 Kinds of groupby attributes and methods	397
40.2 Finding all available attributes and methods	399
40.3 Calling single aggregation methods	400
40.4 <code>head</code> , <code>tail</code> , and <code>nth</code> groupby methods	401
40.5 Non-aggregating methods	404
40.6 Exercises	406
<b>41 Binning Numeric Columns</b>	<b>409</b>
41.1 Grouping with numeric columns	409
41.2 Binning with <code>pd.cut</code>	410
41.3 Quantile binning with <code>pd.qcut</code>	413
41.4 Grouping with bins	414
41.5 Exercises	415
<b>42 Miscellaneous Grouping Functionality</b>	<b>417</b>
42.1 Grouping by columns not in the DataFrame	417
42.2 Grouping Series and aggregating other columns	419
42.3 Grouping by index levels	420
42.4 Changing the direction of grouping	421
42.5 Exercises	422
<b>43 Create Your Own Data Analysis</b>	<b>423</b>
43.1 Overview	423
43.2 Begin Asking and Answering Questions	424
<b>VII Time Series</b>	<b>425</b>
<b>44 Datetime, Timedelta, and Period Objects</b>	<b>427</b>
44.1 Definitions	427
44.2 Date vs Time vs Datetime	427
44.3 Creating single datetime objects in pandas	428
44.4 Timestamp attributes and methods	432
44.5 Creating single timedelta objects in pandas	434
44.6 Timedelta attributes and methods	435
44.7 Creating timedeltas by subtracting datetimes	436
44.8 Creating Period Objects in Pandas	437
44.9 Creating multiple datetimes and timestamps	438
44.10 Exercises	438
<b>45 Selecting Time Series Data</b>	<b>441</b>
45.1 Set the datetime column as the index	441
45.2 Easy subset selection with a <code>DateTimeIndex</code>	442
45.3 Selecting rows at specific frequencies	443
45.4 Upsampling - Increasing the number of rows	445
45.5 Use integers in the offset alias	446
45.6 Exercises	448
<b>46 Grouping by Time</b>	<b>451</b>
46.1 Grouping with the <code>resample</code> method	451
46.2 Grouping by different time periods	453
46.3 Grouping by more than one consecutive offset alias period	455

46.4	Grouping by time with the <code>groupby</code> method	456
46.5	Calling <code>resample</code> on a datetime column	458
46.6	Calling <code>resample</code> on a Series	458
46.7	Exercises	459
<b>47</b>	<b>Rolling Windows</b>	<b>463</b>
47.1	The <code>rolling</code> method	463
47.2	Rolling with offset aliases	465
47.3	DataFrame <code>rolling</code> method	467
47.4	Exercises	468
<b>48</b>	<b>Grouping by Time and another Column</b>	<b>471</b>
48.1	Grouping by an amount of time and another column	472
48.2	Group independently	472
48.3	Using a pivot table with <code>Grouper</code> for easier comparisons	473
48.4	Rolling windows within a group	473
48.5	Exercises	475
<b>49</b>	<b>More Time Series Functionality</b>	<b>477</b>
49.1	Selecting multiple rows at specific frequencies	477
49.2	Shifting the data	481
49.3	Creating date ranges	485
49.4	Exercises	486
<b>VIII</b>	<b>Regular Expressions</b>	<b>489</b>
<b>50</b>	<b>Introduction to Regular Expressions</b>	<b>491</b>
50.1	Regular expressions in Python	491
50.2	Mini-Programming Language	491
50.3	Regular expressions in pandas	492
50.4	Matching with only literal characters	492
50.5	Special Characters	494
50.6	The dot metacharacter	494
50.7	Using raw python strings for regexes	494
50.8	The caret metacharacter <code>^</code>	495
50.9	The dollar sign metacharacter <code>\$</code>	496
50.10	Combining special characters	496
50.11	Setting regex options with flags	496
50.12	Exercises	497
<b>51</b>	<b>Quantifiers</b>	<b>499</b>
51.1	The asterisk metacharacter <code>*</code>	499
51.2	The plus sign metacharacter <code>+</code>	500
51.3	The question mark metacharacter <code>?</code>	500
51.4	The curly braces metacharacter <code>{m,n}</code>	501
51.5	Exercises	501
<b>52</b>	<b>Or Conditions and Character Classes</b>	<b>505</b>
52.1	The pipe metacharacter <code> </code>	505
52.2	The brackets metacharacter <code>[ ]</code>	507
52.3	Character classes with the backslash metacharacter <code>\</code>	510
52.4	Word boundaries with <code>\b</code>	511

52.5	More methods that accept regexes	512
52.6	Exercises	514
<b>53</b>	<b>Grouping and Capturing</b>	<b>517</b>
53.1	Grouping with parentheses ( )	517
53.2	Using parentheses to change operator precedence	518
53.3	Using capture groups with the <code>extract</code> string method	518
53.4	Greedy vs lazy quantifiers	521
53.5	Special syntax for parentheses	522
53.6	Referencing previous groups	522
53.7	Lookaheads and lookbehinds	524
53.8	Positive lookahead assertion - <code>(?=...)</code>	524
53.9	Logical AND operator with positive lookaheads	525
53.10	Negative lookahead assertion - <code>(?!...)</code>	526
53.11	Positive lookbehind assertion - <code>(?&lt;=...)</code>	527
53.12	Negative lookbehind assertion - <code>(?&lt;!...)</code>	527
53.13	Exercises	528
<b>54</b>	<b>Multiline Regex Patterns</b>	<b>531</b>
54.1	Anchors with the multiline flag	532
54.2	Matching every character with	532
54.3	Using multiple flags	533
54.4	Exercises	534
<b>55</b>	<b>Project - Feature Engineering on the Titanic</b>	<b>537</b>
55.1	Exercises	537
<b>IX</b>	<b>Tidy Data</b>	<b>541</b>
<b>56</b>	<b>Tidy Data with <code>melt</code></b>	<b>543</b>
56.1	Tidy data and Lego	543
56.2	Tidy Data	543
56.3	Melting	544
56.4	Exercises	546
<b>57</b>	<b>Reshaping by Pivoting</b>	<b>549</b>
57.1	Inverting melted data with <code>pivot</code>	549
57.2	Pivoting with duplicate values	550
57.3	Using <code>pivot_table</code> to aggregate those values	552
57.4	Exercises	552
<b>58</b>	<b>Common Messy Datasets</b>	<b>555</b>
58.1	Most common messy data problems	555
58.2	Multiple variables are stored in one column	555
58.3	Two or more values are stored in the same cell	557
58.4	Variables are stored in both rows and columns	559
58.5	Steps to produce tidy data	562
58.6	Exercises	563
<b>X</b>	<b>Joining Data</b>	<b>565</b>
<b>59</b>	<b>Automatic Index Alignment</b>	<b>567</b>

59.1	Adding two Series - Not as simple as it sounds	567
59.2	Adding together numpy arrays	569
59.3	Operating on two Series with different index values	570
59.4	Adding Series with duplicate values in the index	571
59.5	Arithmetic operations with two DataFrames	573
59.6	Appending new columns to a DataFrame from a Series	575
59.7	Arithmetic operations with one DataFrame and one Series	578
59.8	DataFrame arithmetic and comparison methods	579
59.9	Exercises	581
<b>60</b>	<b>Concatenating Data</b>	<b>583</b>
60.1	Concatenation with the <code>pd.concat</code> function	583
60.2	Automatic alignment of index always takes place	585
60.3	Change the direction of concatenation with <code>axis</code>	586
60.4	Concatenating many DataFrames together	586
60.5	Concatenating Series together	588
60.6	Appending rows to DataFrames	589
60.7	Adding new columns with <code>assign</code>	591
60.8	Exercises	593
<b>61</b>	<b>Joining DataFrames</b>	<b>595</b>
61.1	Comparing the <code>merge</code> method to <code>pd.concat</code>	595
61.2	Left and right tables	595
61.3	Inner join	596
61.4	Left join	597
61.5	Right join	597
61.6	Outer join	598
61.7	Cross join	598
61.8	Multiple matches per row	599
61.9	Joining on multiple columns	600
61.10	Joining tables from SQL databases	601
61.11	Joining multiple tables together	602
61.12	Further analysis after join	604
61.13	Exercises	604
<b>XI</b>	<b>Fundamentals of SQL</b>	<b>607</b>
<b>62</b>	<b>Intro to Databases and SQL</b>	<b>609</b>
62.1	Databases	609
62.2	SQL	609
62.3	Download DbSchema	610
62.4	The database diagram	610
62.5	Crow's foot notation	613
62.6	Viewing the data	614
62.7	SQL statements	614
62.8	Data types and missing values	615
62.9	Exercises	615
<b>63</b>	<b>The SELECT Statement</b>	<b>617</b>
63.1	Writing and executing SQL statements	617
63.2	Connecting to a database programmatically	618
63.3	SELECT statement clauses	619

63.4	The FROM clause	620
63.5	The LIMIT clause	623
63.6	Database terminology - rows/records and columns/fields	624
63.7	The WHERE clause	624
63.8	The ORDER BY clause	628
63.9	Functions	630
63.10	Arithmetic and comparison operations	632
63.11	Exercises	634
<b>64</b>	<b>GROUP BY and JOIN Clauses</b>	<b>637</b>
64.1	The GROUP BY clause	637
64.2	Ordering after grouping	642
64.3	The JOIN subclause	644
64.4	Right and Full joins	650
64.5	Exercises	650
<b>XII</b>	<b>Visualization with Matplotlib</b>	<b>653</b>
<b>65</b>	<b>Introduction to matplotlib</b>	<b>655</b>
65.1	Two interfaces of matplotlib	655
65.2	Figure - Axes Hierarchy	656
65.3	Setting the size of the figure upon creation	659
65.4	Axes methods	660
65.5	Change tick label and tick line properties with <code>tick_params</code>	667
65.6	Setting multiple properties at the same time with <code>set</code>	668
65.7	Exercises	669
<b>66</b>	<b>Matplotlib Text and Lines</b>	<b>671</b>
66.1	The axes <code>text</code> method	671
66.2	Creating horizontal lines with <code>hlines</code>	677
66.3	Create vertical lines with <code>vlines</code>	680
66.4	Add grid lines with the <code>grid</code> method	681
66.5	Aligning text horizontally and vertically	683
66.6	Add text with arrows using the <code>annotate</code> method	685
66.7	Exercises	689
<b>67</b>	<b>Matplotlib Resolution</b>	<b>691</b>
67.1	Matplotlib inches	691
67.2	Creating figures with custom DPI	694
67.3	Text and line “points”	695
67.4	Run configuration settings	696
67.5	Creating style sheets	700
67.6	Exercises	703
<b>68</b>	<b>Matplotlib Patches and Colors</b>	<b>705</b>
68.1	Adding matplotlib patches	705
68.2	Circle patches	705
68.3	Ellipse patches	707
68.4	Rectangle patches	708
68.5	Polygon patches	708
68.6	Arc patches	709
68.7	Wedge patches	710

68.8 Matplotlib colors . . . . .	712
68.9 Color transparency . . . . .	717
68.10 Layering with zorder . . . . .	718
68.11 Gray scale . . . . .	719
68.12 Colormaps . . . . .	720
68.13 Filling between two lines . . . . .	722
68.14 Creating a basketball court . . . . .	724
68.15 Exercises . . . . .	728
<b>69 Matplotlib Line Plots</b>	<b>731</b>
69.1 Axes API . . . . .	731
69.2 Line plots with the <code>plot</code> method . . . . .	731
69.3 Integration with pandas . . . . .	734
69.4 Color cycle . . . . .	738
69.5 More line plots . . . . .	739
69.6 Adding a legend . . . . .	742
69.7 Exercises . . . . .	747
<b>70 Matplotlib Scatter and Bar Plots</b>	<b>749</b>
70.1 Scatter plots . . . . .	749
70.2 Change scatter plot point size . . . . .	759
70.3 Bar plots . . . . .	762
70.4 Exercises . . . . .	770
<b>71 Matplotlib Distribution Plots</b>	<b>771</b>
71.1 Histograms . . . . .	772
71.2 Box and whisker plots . . . . .	778
71.3 Exercises . . . . .	780
<b>72 Best of the Rest of Matplotlib</b>	<b>781</b>
72.1 Axes spines . . . . .	781
72.2 The <code>xaxis</code> and <code>yaxis</code> objects . . . . .	782
72.3 Tick locators . . . . .	784
72.4 Tick formatters . . . . .	785
72.5 Minor ticks . . . . .	785
72.6 Horizontal and vertical lines that span the axes . . . . .	787
72.7 Plotting with dates . . . . .	788
72.8 Using a different scale for the axis . . . . .	791
72.9 Adding images . . . . .	794
72.10 Coordinate systems . . . . .	799
72.11 Figure methods . . . . .	803
72.12 Creating a grid of axes . . . . .	806
72.13 Exercises . . . . .	808
<b>XIII Visualization with Pandas and Seaborn</b>	<b>809</b>
<b>73 Plotting with pandas Series</b>	<b>811</b>
73.1 Line plots . . . . .	812
73.2 Bar plots . . . . .	814
73.3 Distribution plots . . . . .	815
73.4 Pie Charts . . . . .	816
73.5 Area Plots . . . . .	817

73.6 Adding a plot to a previously made axes . . . . .	817
<b>74 Plotting with pandas DataFrames</b>	<b>819</b>
74.1 Line plots . . . . .	819
74.2 Bar plots . . . . .	821
74.3 Plotting on separate axes . . . . .	824
74.4 Distribution plots . . . . .	825
74.5 Scatter and Hexbin . . . . .	826
74.6 Area plots . . . . .	830
<b>75 Seaborn Axes Plots</b>	<b>835</b>
75.1 The seaborn API . . . . .	835
75.2 seaborn integration with pandas . . . . .	836
75.3 Distribution Plots . . . . .	836
75.4 Seaborn style sheets . . . . .	841
75.5 Other distribution plots . . . . .	841
75.6 Automatic grouping by category . . . . .	842
75.7 Grouping within groups with <code>hue</code> . . . . .	844
75.8 Tidy data . . . . .	845
75.9 Grouping and Aggregating Plots . . . . .	846
75.10 Raw data plots . . . . .	853
75.11 Scatter plots with linear regression lines using <code>regplot</code> . . . . .	858
75.12 Ordered categorical data . . . . .	860
75.13 Exercises . . . . .	862
<b>76 Seaborn Grid Plots</b>	<b>863</b>
76.1 Grids by categories . . . . .	863
76.2 Scatter and line plot grids . . . . .	867
76.3 Regression grid plots . . . . .	868
76.4 Bivariate distributions grids . . . . .	869
76.5 Scatter plot grids of multiple column combinations . . . . .	870
76.6 Hierarchical cluster map . . . . .	871