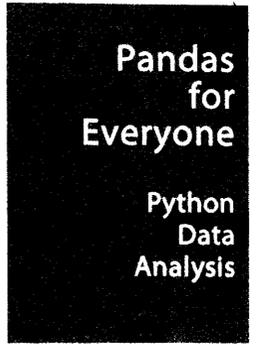


C O N T E N T S



口絵.....	i
序文.....	xix
まえがき.....	xx
本書の構成.....	xxii
本書の読み方.....	xxiv
データの入手方法など.....	xxv
謝辞.....	xxvii

第 1 部 基本的な使い方を学ぶ

第 1 章 DataFrame の基礎.....	2
1.1 はじめに.....	2
1.2 最初のデータセットをロードする.....	3
1.3 列、行、セルを見る.....	6
1.3.1 列を絞り込む.....	7
1.3.2 行を絞り込む.....	8
1.3.3 組み合わせて絞り込む.....	12

1.4 グループ化と集約.....	18
1.4.1 グループごとの平均値.....	19
1.4.2 グループごとの度数 / 頻度.....	24
1.5 基本的なグラフ.....	24
1.6 まとめ.....	25
第2章 pandas のデータ構造.....	26
2.1 はじめに.....	26
2.2 データを自作する.....	27
2.2.1 Series を作る.....	27
2.2.2 DataFrame を作る.....	28
2.3 Series について.....	30
2.3.1 Series は ndarray に似たもの.....	32
2.3.2 真偽値による絞り込み.....	33
2.3.3 演算の自動的な整列とベクトル化 (ブロードキャスト).....	35
2.4 DataFrame について.....	38
2.4.1 真偽値による絞り込み : DataFrame.....	38
2.4.2 演算による整列とベクトル化 (ブロードキャスト).....	39
2.5 Series と DataFrame の書き換え.....	41
2.5.1 列を追加する.....	41
2.5.2 列を直接変更する.....	42
2.5.3 列を捨てる.....	45
2.6 データのエクスポートとインポート.....	46
2.6.1 pickle.....	46
2.6.2 CSV.....	48
2.6.3 Excel.....	49
2.6.4 feather フォーマット : R 言語とのインターフェイス.....	50
2.6.5 その他のデータ出力形式.....	51
2.7 まとめ.....	51

第3章 プロットによるグラフ描画.....	52
3.1 はじめに.....	52
3.2 matplotlib.....	53
3.3 matplotlib による統計的グラフィックス.....	60
3.3.1 1 変量データ.....	60
3.3.2 2 変量データ.....	61
3.3.3 多変量データ.....	63
3.4 seaborn.....	65
3.4.1 1 変量データ.....	66
3.4.2 2 変量データ.....	70
3.4.3 多変量データ.....	80
3.5 pandas のオブジェクト.....	90
3.5.1 ヒストグラム.....	90
3.5.2 密度プロット.....	92
3.5.3 散布図.....	92
3.5.4 hexbin プロット.....	93
3.5.5 箱ひげ図.....	94
3.6 seaborn のテーマとスタイル.....	95
3.7 まとめ.....	98

第2部 データ操作によるクリーニング

第4章 データを組み立てる.....	100
4.1 はじめに.....	100
4.2 " 整然データ ".....	101
4.2.1 データセットを組み合わせる.....	101
4.3 連結.....	101
4.3.1 行の追加.....	102
4.3.2 列の追加.....	106
4.3.3 インデックスが異なる連結.....	107

4.4 複数のデータセットをマージする.....	110
4.4.1 1対1のマージ.....	112
4.4.2 多対1のマージ.....	112
4.4.3 多対多のマージ.....	113
4.5 まとめ.....	115
第5章 欠損データへの対応.....	116
5.1 はじめに.....	116
5.2 NaNとは何か.....	117
5.3 欠損値はどこから来るのか.....	118
5.3.1 データのロード.....	118
5.3.2 マージされたデータ.....	120
5.3.3 ユーザー入力.....	121
5.3.4 インデックスの振り直し.....	122
5.4 欠損データの扱い.....	124
5.4.1 欠損データを数える.....	124
5.4.2 欠損データのクリーニング.....	125
5.4.3 欠損データとの計算.....	128
5.5 まとめ.....	129
第6章 "整然データ"を作る.....	130
6.1 はじめに.....	130
6.2 複数列に(変数ではなく)値が入っているとき.....	131
6.2.1 1列に集める.....	131
6.2.2 複数の列を残す.....	133
6.3 複数の変数を含む列がある場合.....	135
6.3.1 列を分割して追加する単純な方法.....	136
6.3.2 分割と結合を一度に行う(単純な方法).....	138
6.3.3 分割と結合を一度に行う(より複雑な方法).....	138
6.4 行と列の両方に変数があるとき.....	140
6.5 1個の表に観察単位が複数あるとき(正規化).....	142

6.6 同じ観察単位が複数の表にまたがっているとき.....	144
6.6.1 ループを使って複数のファイルをロードする.....	146
6.6.2 リスト内包処理を使って複数のファイルをロードする.....	148
6.7 まとめ.....	148

第3部 データの準備—変換／整形／結合など

第7章 データ型の概要と変換.....	150
7.1 はじめに.....	150
7.2 データ型.....	151
7.3 型変換.....	151
7.3.1 文字列オブジェクトへの変換.....	152
7.3.2 数値への変換.....	152
7.4 カテゴリ型データ.....	157
7.4.1 カテゴリ型への変換.....	158
7.4.2 カテゴリ型データを操作する.....	159
7.5 まとめ.....	159
第8章 テキスト文字列の操作.....	160
8.1 はじめに.....	160
8.2 文字列.....	161
8.2.1 文字列の抽出とスライス.....	161
8.2.2 文字列の最後の文字を取得する.....	163
8.3 文字列メソッド.....	164
8.4 その他の文字列メソッド.....	166
8.4.1 join.....	166
8.4.2 splitlines.....	166
8.5 文字列のフォーマット.....	167
8.5.1 フォーマットの形式.....	168
8.5.2 文字列の書式化.....	168

8.5.3	数値の書式化.....	169	10.2.3	集約関数.....	201
8.5.4	Cのprintfスタイルによる書式化.....	170	10.2.4	複数の関数を同時に計算する.....	203
8.5.5	フォーマット済み文字列リテラル (Python 3.6 から).....	170	10.2.5	agg/aggregateでdictを使う.....	204
8.6	正規表現.....	171	10.3	変換 (transform).....	205
8.6.1	パターンとのマッチ.....	172	10.3.1	標準スコアの例.....	205
8.6.2	パターンを見つける.....	175	10.4	フィルタリング.....	210
8.6.3	パターンを置換する.....	175	10.5	DataFrameGroupBy オブジェクト.....	211
8.6.4	パターンをコンパイルする.....	176	10.5.1	グループ.....	211
8.7	regex ライブラリ.....	177	10.5.2	複数の変数に関わるグループ計算.....	212
8.8	まとめ.....	177	10.5.3	グループの抽出.....	213
第 9 章	apply による関数の適用.....	178	10.5.4	グループごとの反復処理.....	213
9.1	はじめに.....	178	10.5.5	複数変数のグループ.....	215
9.2	関数.....	179	10.5.6	結果を平坦化する.....	216
9.3	apply の基本.....	180	10.6	マルチインデックスを使う.....	217
9.3.1	Series に適用する.....	180	10.7	まとめ.....	221
9.3.2	DataFrame に適用する.....	182	第 11 章	日付/時刻データの操作.....	222
9.4	apply の応用.....	185	11.1	はじめに.....	222
9.4.1	列ごとの演算.....	187	11.2	Python の datetime オブジェクト.....	223
9.4.2	行ごとの演算.....	189	11.3	datetime への変換.....	223
9.5	関数のベクトル化.....	191	11.4	日付を含むデータをロードする.....	226
9.5.1	NumPy を使ったベクトル化.....	192	11.5	日付のコンポーネントを抽出する.....	227
9.5.2	numba を使ったベクトル化.....	193	11.6	日付の計算と timedelta.....	229
9.6	ラムダ関数.....	194	11.7	datetime のメソッド.....	231
9.7	まとめ.....	196	11.8	株価データを取得する.....	233
第 10 章	groupby 演算による分割 - 適用 - 結合.....	197	11.9	日付によるデータの絞り込み.....	234
10.1	はじめに.....	197	11.9.1	DatetimeIndex オブジェクト.....	235
10.2	集約.....	198	11.9.2	TimedeltaIndex オブジェクト.....	236
10.2.1	1 個の変数で分割する基本的な集約.....	198	11.10	日付の範囲.....	236
10.2.2	組み込みの集約メソッド.....	200	11.10.1	周期.....	238
			11.10.2	オフセット.....	239

11.11	値をシフトする.....	240
11.12	リサンプリング.....	246
11.13	時間帯.....	247
11.14	まとめ.....	249

第4部 モデルをデータに適合させる

第12章	線形モデル.....	252
12.1	はじめに.....	252
12.2	単純な線形回帰.....	252
12.2.1	Python 統計ライブラリ statsmodels を使う.....	253
12.2.2	Python 機械学習ライブラリ sklearn を使う.....	255
12.3	重回帰.....	257
12.3.1	Python 統計ライブラリ statsmodels を使う.....	257
12.3.2	statsmodels でカテゴリ変数を使う.....	258
12.3.3	Python 機械学習ライブラリ sklearn を使う.....	260
12.3.4	sklearn でカテゴリ変数を使う.....	261
12.4	sklearn でインデックスラベルを残す.....	262
12.5	まとめ.....	263
第13章	一般化線形モデル.....	264
13.1	はじめに.....	264
13.2	ロジスティック回帰.....	264
13.2.1	Python 統計ライブラリ statsmodels を使う.....	266
13.2.2	Python 機械学習ライブラリ sklearn を使う.....	268
13.3	ポアソン回帰.....	269
13.3.1	statsmodels の poisson 関数.....	269
13.3.2	過分散のための「負の2項回帰」.....	271
13.4	その他の一般化線形モデル.....	272

13.5	生存分析.....	272
13.5.1	Cox モデルの前提をチェックする.....	275
13.6	まとめ.....	277

第14章	モデルを診断する.....	278
14.1	はじめに.....	278
14.2	残差.....	278
14.2.1	Q-Q プロット.....	281
14.3	複数のモデルを比較する.....	283
14.3.1	線形モデルの比較.....	283
14.3.2	GLM モデルの比較.....	286
14.4	k 分割交差検証.....	288
14.5	まとめ.....	292
第15章	正則化で過学習に対処する.....	293
15.1	はじめに.....	293
15.2	なぜ正則化するのか.....	293
15.3	LASSO 回帰.....	295
15.4	リッジ回帰.....	297
15.5	Elastic Net.....	298
15.6	交差検証.....	300
15.7	まとめ.....	303
第16章	クラスタリング.....	304
16.1	はじめに.....	304
16.2	k 平均法.....	304
16.2.1	主成分分析で次元を減らす.....	307
16.3	階層的クラスタリング.....	311
16.3.1	完全リンク法.....	312
16.3.2	単一リンク法.....	312

16.3.3 群平均法.....	313
16.3.4 重心法.....	314
16.3.5 色分けの「しきい値」を設定する.....	314
16.4 まとめ.....	315

第5部 締めくくり—次のステップへ

第17章 pandas 周辺の強力な機能.....	318
17.1 Python の科学計算スタック.....	318
17.2 コードの性能.....	319
17.2.1 実行速度を計測する.....	319
17.2.2 プロファイリングを行う.....	321
17.3 大きなデータをより速く処理する.....	321
第18章 さらなる学びのための情報源.....	322
18.1 1人歩きは危険だ！.....	322
18.2 地元でのミートアップ.....	322
18.3 カンファレンス.....	323
18.4 インターネット.....	324
18.5 ポッドキャスト.....	324
18.6 まとめ.....	325

第6部 付録

付録A インストール.....	328
付録B コマンドライン.....	330
付録C プロジェクトのテンプレート.....	332
付録D Python の使い方.....	333

付録E ワーキングディレクトリ.....	336
付録F 環境.....	338
付録G パッケージのインストール.....	341
付録H ライブラリのインポート.....	343
付録I リスト.....	345
付録J タプル.....	347
付録K 辞書.....	348
付録L 値のスライス.....	351
付録M ループ.....	353
付録N 内包表記 (comprehension).....	355
付録O 関数.....	357
付録P 範囲とジェネレータ.....	362
付録Q 複数代入.....	365
付録R numpy の ndarray.....	367
付録S クラス.....	369
付録T Odo (The Shapeshifter).....	371
参考文献.....	372
索引.....	373