

Python para Análise de Dados

TRATAMENTO DE DADOS COM PANDAS, NUMPY E IPYTHON

Wes McKinney

Novatec

Authorized Portuguese translation of the English edition of Python for Data Analysis, 2E, ISBN 9781491957660
© 2018 William Wesley McKinney. This translation is published and sold by permission of O'Reilly Media, Inc.,
the owner of all rights to publish and sell the same.

Tradução em português autorizada da edição em inglês da obra Python for Data Analysis, 2E, ISBN 9781491957660
© 2018 William Wesley McKinney. Esta tradução é publicada e vendida com a permissão da O'Reilly Media,
Inc., detentora de todos os direitos para publicação e venda desta obra.

Copyright © 2018 da Novatec Editora Ltda.

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998. É proibida a reprodução desta obra,
mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e da Editora.

Editor: Rubens Prates

Tradução: Lúcia A. Kinoshita

Revisão gramatical: Tássia Carvalho

Editoração eletrônica: Carolina Kuwabata

Capa: Carolina Kuwabata

ISBN: 978-85-7522-647-6

Histórico de impressões:

Janeiro/2018 Primeira edição

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110

02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

Email: novatec@novatec.com.br

Site: www.novatec.com.br

Twitter: twitter.com/novateceditora

Facebook: facebook.com/novatec

LinkedIn: linkedin.com/in/novatec

Sumário

Prefácio	13
Capítulo 1 ■ Informações preliminares	19
1.1 De que se trata este livro?	19
Quais tipos de dados?	19
1.2 Por que Python para análise de dados?.....	20
Python como aglutinador	21
Resolvendo o problema de “duas linguagens”	21
Por que não Python?.....	21
1.3 Bibliotecas Python essenciais.....	22
NumPy	22
pandas.....	23
matplotlib	25
IPython e Jupyter	25
SciPy	26
scikit-learn.....	27
statsmodels.....	27
1.4 Instalação e configuração	28
Windows	28
Apple (OS X, MacOS)	29
GNU/Linux.....	29
Instalando ou atualizando pacotes Python.....	30
Python 2 e Python 3.....	31
Ambientes de desenvolvimento integrado (IDEs) e editores de texto	31
1.5 Comunidade e conferências.....	32
1.6 Navegando pelo livro	33
Exemplos de código	34
Dados para os exemplos	34
Convenções de importação.....	35
Jargão	35

Capítulo 2 = Básico da linguagem Python, IPython e notebooks Jupyter36

2.1 Interpretador Python.....	37
2.2 Básico sobre o IPython.....	39
Executando o shell IPython.....	39
Executando o notebook Jupyter.....	40
Preenchimento automático com tabulação	43
Introspecção	45
Comando %run	47
Executando código da área de transferência.....	48
Atalhos de teclado no terminal	49
Sobre os comandos mágicos	50
Integração com a matplotlib	53
2.3 Básico da linguagem Python	53
Semântica da linguagem.....	54
Tipos escalares.....	63
Controle de fluxo	72

Capítulo 3 = Estruturas de dados embutidas, funções e arquivos78

3.1 Estruturas de dados e sequências.....	78
Tupla	78
Lista	82
Funções embutidas para sequências.....	88
List, set e dict comprehensions	97
3.2 Funções.....	100
Namespaces, escopo e funções locais	101
Devolvendo diversos valores	102
Funções são objetos	103
Funções anônimas (lambdas)	105
Currying: aplicação parcial dos argumentos	106
Geradores	107
Erros e tratamento de exceção	110
3.3 Arquivos e o sistema operacional.....	113
Bytes e Unicode com arquivos	117
3.4 Conclusão.....	119

Capítulo 4 = Básico sobre o NumPy: arrays e processamento vetorizado 120

4.1 O ndarray do NumPy: um objeto array multidimensional	122
Criando ndarrays.....	124
Tipos de dados para ndarrays.....	126
Aritmética com arrays NumPy	129
Indexação básica e fatiamento	131
Indexação booleana	136

Indexação sofisticada	140
Transposição de arrays e troca de eixos.....	142
4.2 Funções universais: funções rápidas de arrays para todos os elementos	144
4.3 Programação orientada a arrays.....	147
Expressando uma lógica condicional como operações de array.....	149
Métodos matemáticos e estatísticos	150
Métodos para arrays booleanos	152
Ordenação	153
Unicidade e outras lógicas de conjuntos.....	154
4.4 Entrada e saída de arquivos com arrays.....	155
4.5 Álgebra linear	156
4.6 Geração de números pseudoaleatórios	159
4.7 Exemplo: passeios aleatórios	161
Simulando vários passeios aleatórios de uma só vez	163
4.8 Conclusão.....	164
Capítulo 5 = Introdução ao pandas.....	165
5.1 Introdução às estruturas de dados do pandas.....	166
Series.....	166
DataFrame.....	171
Objetos Index	179
5.2 Funcionalidades essenciais.....	181
Reindexação	181
Descartando entradas de um eixo.....	184
Indexação, seleção e filtragem.....	186
Índices inteiros.....	192
Aritmética e alinhamento de dados.....	193
Aplicação de funções e mapeamento.....	200
Ordenação e classificação.....	203
Índices de eixos com rótulos duplicados	207
5.3 Resumindo e calculando estatísticas descritivas.....	208
Correlação e covariância	212
Valores únicos, contadores de valores e pertinência.....	214
5.4 Conclusão.....	217
Capítulo 6 = Carga de dados, armazenagem e formatos de arquivo	218
6.1 Lendo e escrevendo dados em formato-texto	218
Lendo arquivos-texto em partes	226
Escrevendo dados em formato-texto	228
Trabalhando com formatos delimitados.....	230
Dados JSON	232
XML e HTML: web scraping.....	234

6.2 Formatos de dados binários	239
Usando o formato HDF5	240
Lendo arquivos do Microsoft Excel.....	242
6.3 Interagindo com APIs web	244
6.4 Interagindo com bancos de dados.....	245
6.5 Conclusão	247
Capítulo 7 = Limpeza e preparação dos dados.....	248
7.1 Tratando dados ausentes	248
Filtrando dados ausentes	250
Preenchendo dados ausentes	253
7.2 Transformação de dados.....	256
Removendo duplicatas	256
Transformando dados usando uma função ou um mapeamento	257
Substituindo valores.....	260
Renomeando os índices dos eixos.....	261
Discretização e compartimentalização (binning).....	263
Detectando e filtrando valores discrepantes	266
Permutação e amostragem aleatória	268
Calculando variáveis indicadoras/dummy.....	269
7.3 Manipulação de strings	274
Métodos de objetos string	274
Expressões regulares.....	276
Funções de string vetorizadas no pandas	280
7.4 Conclusão	283
Capítulo 8 = Tratamento de dados: junção, combinação e reformatação	284
8.1 Indexação hierárquica	284
Reorganizando e ordenando níveis	288
Estatísticas de resumo por nível.....	289
Indexando com as colunas de um DataFrame	290
8.2 Combinando e mesclando conjuntos de dados.....	291
Junções no DataFrame no estilo de bancos de dados	292
Fazendo merge com base no índice	298
Concatenando ao longo de um eixo	303
Combinando dados com sobreposição.....	309
8.3 Reformatação e pivoteamento	311
Reformatação com indexação hierárquica.....	311
Fazendo o pivoteamento de um formato “longo” para um formato “largo”	315
Pivoteamento do formato “largo” para o formato “longo”	319
8.4 Conclusão.....	321

Capítulo 9 = Plotagem e visualização	322
9.1 Introdução rápida à API da matplotlib	323
Figuras e subplotagens	324
Cores, marcadores e estilos de linha	328
Tiques, rótulos e legendas	330
Anotações e desenhos em uma subplotagem	334
Salvando plotagens em arquivos	336
Configuração da matplotlib.....	337
9.2 Plottagem com o pandas e o seaborn	338
Plotagens de linha	338
Plotagem de barras.....	341
Histogramas e plotagens de densidade.....	346
Plotagens de dispersão ou de pontos	348
Grades de faceta e dados de categoria	350
9.3 Outras ferramentas de visualização de Python.....	352
9.4 Conclusão.....	353
Capítulo 10 = Agregação de dados e operações em grupos	354
10.1 Funcionamento de GroupBy	355
Iterando por grupos	359
Selecionando uma coluna ou um subconjunto de colunas.....	361
Agrupando com dicionários e Series	362
Agrupando com funções	363
Agrupando por níveis de índice	364
10.2 Agregação de dados.....	365
Aplicação de função nas colunas e aplicação de várias funções	367
Devolvendo dados agregados sem índices de linha	371
10.3 Método apply: separar-aplicar-combinar genérico.....	372
Suprimindo as chaves de grupo	375
Análise de quantis e de buckets	376
Exemplo: preenchendo valores ausentes com valores específicos de grupo	377
Exemplo: amostragem aleatória e permutação.....	380
Exemplo: média ponderada de grupos e correlação	382
Exemplo: regressão linear nos grupos	385
10.4 Tabelas pivôs e tabulação cruzada.....	386
Tabulações cruzadas: crosstab.....	389
10.5 Conclusão.....	390
Capítulo 11 = Séries temporais.....	391
11.1 Tipos de dados e ferramentas para data e hora	392
Conversão entre string e datetime.....	393
11.2 Básico sobre séries temporais.....	396
Indexação, seleção e geração de subconjuntos	398

Séries temporais com índices duplicados.....	402
11.3 Intervalos de datas, frequências e deslocamentos	403
Gerando intervalos de datas	404
Frequências e offset de datas	407
Deslocamento de datas (adiantando e atrasando)	409
11.4 Tratamento de fusos horários.....	413
Localização e conversão dos fusos horários.....	414
Operações com objetos Timestamp que consideram fusos horários	417
Operações entre fusos horários diferentes.....	418
11.5 Períodos e aritmética com períodos.....	419
Conversão de frequência de períodos.....	420
Frequências de período trimestrais	422
Convertendo timestamps para períodos (e vice-versa)	424
Criando um PeriodIndex a partir de arrays	426
11.6 Reamostragem e conversão de frequências	429
Downsampling	431
Upsampling e interpolação	434
Reamostragem com períodos.....	436
11.7 Funções de janela móvel	438
Funções exponencialmente ponderadas	441
Funções de janela móvel binárias.....	442
Funções de janela móvel definidas pelo usuário	444
11.8 Conclusão.....	445
Capítulo 12 = Pandas avançado.....	446
12.1 Dados categorizados	446
Informações básicas e motivação	446
Tipo Categorical do pandas.....	448
Processamentos com Categoricals.....	451
Métodos para dados categorizados	454
12.2 Uso avançado de GroupBy	458
Transformações de grupos e GroupBys “não encapsulados”	458
Reamostragem de tempo em grupos	463
12.3 Técnicas para encadeamento de métodos.....	465
Método pipe	466
12.4 Conclusão.....	468
Capítulo 13 = Introdução às bibliotecas de modelagem em Python	469
13.1 Interface entre o pandas e o código dos modelos	470
13.2 Criando descrições de modelos com o Patsy	473
Transformações de dados em fórmulas do Patsy.....	476
Dados categorizados e o Patsy	478

13.3 Introdução ao statsmodels.....	482
Estimando modelos lineares.....	482
Estimando processos de séries temporais.....	486
13.4 Introdução ao scikit-learn.....	487
13.5 Dando prosseguimento à sua educação.....	492
Capítulo 14 = Exemplos de análises de dados.....	493
14.1 Dados de 1.USA.gov do Bitly.....	493
Contando fusos horários em Python puro.....	494
Contando fusos horários com o pandas.....	497
14.2 Conjunto de dados do MovieLens 1M.....	505
Avaliando a discrepância nas avaliações.....	511
14.3 Nomes e bebês americanos de 1880 a 2010.....	513
Analisando tendências para os nomes.....	519
14.4 Banco de dados de alimentos do USDA.....	529
14.5 Banco de dados da Federal Election Commission em 2012.....	536
Estatísticas sobre as doações de acordo com a profissão e o empregador.....	540
Separando os valores das doações em buckets.....	543
Estatísticas sobre as doações conforme o estado.....	546
14.6 Conclusão.....	547
Apêndice A = NumPy avançado.....	548
A.1 Organização interna do objeto ndarray.....	548
A hierarquia de dtypes do NumPy.....	549
A.2 Manipulação avançada de arrays.....	551
Redefinindo o formato de arrays.....	551
Ordem C versus ordem Fortran.....	553
Concatenando e separando arrays.....	554
Repetindo elementos: tile e repeat.....	557
Equivalentes à indexação sofisticada: take e put.....	559
A.3 Broadcasting.....	561
Broadcasting em outros eixos.....	563
Definindo valores de array para broadcasting.....	566
A.4 Usos avançados de ufuncs.....	567
Métodos de instância de ufuncs.....	567
Escrevendo novas ufuncs em Python.....	570
A.5 Arrays estruturados e de registros.....	571
dtypes aninhados e campos multidimensionais.....	572
Por que usar arrays estruturados?.....	573
A.6 Mais sobre ordenação.....	573
Ordenações indiretas: argsort e lexsort.....	575
Algoritmos de ordenação alternativos.....	577

Ordenando arrays parcialmente.....	578
numpy.searchsorted: encontrando elementos em um array ordenado.....	579
A.7 Escrevendo funções NumPy rápidas com o Numba.....	580
Criando objetos numpy.ufunc personalizados com o Numba.....	582
A.8 Operações avançadas de entrada e saída com arrays.....	583
Arquivos mapeados em memória.....	583
HDF5 e outras opções para armazenagem de arrays.....	584
A.9 Dicas para o desempenho.....	585
A importância da memória contígua.....	585
Apêndice B = Mais sobre o sistema IPython.....	588
B.1 Usando o histórico de comandos.....	588
Pesquisando e reutilizando o histórico de comandos.....	588
Variáveis de entrada e de saída.....	589
B.2 Interagindo com o sistema operacional.....	590
Comandos do shell e aliases.....	591
Sistema de marcadores de diretórios.....	593
B.3 Ferramentas para desenvolvimento de software.....	593
Depurador interativo.....	594
Medindo o tempo de execução de um código: %time e %timeit.....	599
Geração básica de perfis: %prun e %run -p.....	602
Gerando o perfil de uma função linha a linha.....	604
B.4 Dicas para um desenvolvimento de código produtivo usando o IPython.....	607
Recarregando dependências de módulos.....	607
Dicas para design de código.....	608
B.5 Recursos avançados do IPython.....	610
Deixando suas próprias classes mais apropriadas ao IPython.....	610
Perfis e configuração.....	611
B.6 Conclusão.....	613