

# Análise de dados com Python e Pandas

Daniel Y. Chen

 Pearson  
Novatec

Authorized translation from the English language edition, entitled PANDAS FOR EVERYONE: PYTHON DATA ANALYSIS, 1st Edition by DANIEL CHEN, published by Pearson Education, Inc, publishing as Addison-Wesley Professional, Copyright © 2018 by Pearson Education, Inc.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc. PORTUGUESE language edition published by NOVATEC EDITORA LTDA., Copyright © 2018.

Tradução autorizada da edição original em inglês, intitulada PANDAS FOR EVERYONE: PYTHON DATA ANALYSIS, 1st Edition por DANIEL CHEN, publicada pela Pearson Education, Inc, publicando como Addison-Wesley Professional, Copyright © 2018 por Pearson Education, Inc.

Todos os direitos reservados. Nenhuma parte deste livro pode ser reproduzida ou transmitida por qualquer forma ou meio, eletrônica ou mecânica, incluindo fotocópia, gravação ou qualquer sistema de armazenamento de informação, sem a permissão da Pearson Education, Inc. Edição em Português publicada pela NOVATEC EDITORA LTDA., Copyright © 2018.

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

Design da capa: Chuti Prasertsith

Ilustração da capa: theromb/Shutterstock

ISBN: 978-85-7522-699-5

Histórico de impressões:

Setembro/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](mailto:novatec@novatec.com.br)

Site: [www.novatec.com.br](http://www.novatec.com.br)

Twitter: [twitter.com/novateceditora](https://twitter.com/novateceditora)

Facebook: [facebook.com/novatec](https://facebook.com/novatec)

LinkedIn: [linkedin.com/in/novatec](https://linkedin.com/in/novatec)

# Sumário

<b>Apresentação.....</b>	<b>15</b>
<b>Prefácio .....</b>	<b>17</b>
<b>Agradecimentos.....</b>	<b>25</b>
<b>Sobre o autor .....</b>	<b>29</b>
<b>Parte I = Introdução.....</b>	<b>30</b>
<b>Capítulo 1 = Básico sobre o DataFrame do Pandas.....</b>	<b>31</b>
1.1 Introdução.....	31
1.2 Carregando seu primeiro conjunto de dados .....	32
1.3 Observando colunas, linhas e células .....	36
1.3.1 Obtendo subconjuntos de colunas.....	36
1.3.2 Obtendo subconjuntos de linhas.....	38
1.3.3 Combinando tudo.....	43
1.4 Cálculos agrupados e agregados.....	51
1.4.1 Médias agrupadas.....	52
1.4.2 Contadores de frequência agrupados .....	56
1.5 Plotagem básica.....	57
1.6 Conclusão.....	58
<b>Capítulo 2 = Estruturas de dados do Pandas.....</b>	<b>59</b>
2.1 Introdução .....	59
2.2 Criando seus próprios dados .....	60
2.2.1 Criando uma Series .....	60
2.2.2 Criando um DataFrame .....	61
2.3 Series.....	63
2.3.1 Series é semelhante a ndarray.....	65
2.3.2 Subconjuntos com booleanos: Series.....	66
2.3.3 Operações são alinhadas e vetorizadas automaticamente (broadcasting).....	69

2.4 DataFrame .....	73
2.4.1 Subconjuntos com booleanos: DataFrames .....	73
2.4.2 Operações são alinhadas e vetorizadas automaticamente (broadcasting)74	
2.5 Fazendo alterações em Series e em DataFrames .....	76
2.5.1 Adicionando mais colunas .....	76
2.5.2 Alterando diretamente uma coluna .....	78
2.5.3 Descartando valores .....	81
2.6 Exportando e importando dados .....	82
2.6.1 pickle .....	82
2.6.2 CSV .....	85
2.6.3 Excel .....	86
2.6.4 Formato feather para interface com R .....	87
2.6.5 Outros tipos de saída de dados .....	87
2.7 Conclusão .....	88

### **Capítulo 3 = Introdução à plotagem .....89**

3.1 Introdução .....	89
3.2 Matplotlib .....	91
3.3 Gráficos estatísticos usando a matplotlib .....	97
3.3.1 Univariado .....	98
3.3.2 Bivariado .....	99
3.3.3 Dados multivariados .....	100
3.4 seaborn .....	102
3.4.1 Univariado .....	103
3.4.2 Dados bivariados .....	106
3.4.3 Dados multivariados .....	114
3.5 Objetos do Pandas .....	123
3.5.1 Histogramas .....	123
3.5.2 Plotagem de densidade .....	125
3.5.3 Gráfico de dispersão .....	125
3.5.4 Plotagem hexbin .....	126
3.5.5 Gráfico de caixa .....	127
3.6 Temas e estilos do seaborn .....	127
3.7 Conclusão .....	129

### **Parte II = Manipulação de dados ..... 131**

#### **Capítulo 4 = Preparação dos dados ..... 132**

4.1 Introdução .....	132
4.2 Tidy Data .....	133
4.2.1 Combinando conjuntos de dados .....	133

4.3 Concatenação .....	134
4.3.1 Adicionando linhas.....	134
4.3.2 Adicionando colunas.....	139
4.3.3 Concatenação com índices diferentes .....	140
4.4 Combinando vários conjuntos de dados .....	144
4.4.1 Merge um a um .....	146
4.4.2 Merge de muitos para um.....	147
4.4.3 Merge de muitos para muitos .....	148
4.5 Conclusão .....	150
<b>Capítulo 5 = Dados ausentes .....</b>	<b>151</b>
5.1 Introdução.....	151
5.2 O que é um valor NaN?.....	152
5.3 De onde vêm os valores ausentes?.....	153
5.3.1 Carga de dados .....	153
5.3.2 Dados combinados.....	155
5.3.3 Valores de entrada do usuário .....	157
5.3.4 Reindexação .....	158
5.4 Trabalhando com dados ausentes.....	160
5.4.1 Encontrando e contando dados ausentes .....	160
5.4.2 Limpando dados ausentes .....	162
5.4.3 Cálculos com dados ausentes.....	165
5.5 Conclusão.....	166
<b>Capítulo 6 = Tidy data (dados organizados) .....</b>	<b>167</b>
6.1 Introdução .....	167
6.2 Colunas contêm valores, e não variáveis .....	168
6.2.1 Mantendo uma coluna fixa .....	168
6.2.2 Mantendo várias colunas fixas.....	171
6.3 Colunas contendo diversas variáveis.....	173
6.3.1 Separar e adicionar colunas individualmente (método simples).....	174
6.3.2 Separar e combinar em um único passo (método simples).....	177
6.3.3 Separar e combinar em um único passo (método mais complicado)....	178
6.4 Variáveis tanto em linhas quanto em colunas.....	180
6.5 Várias unidades de observação em uma tabela (normalização) .....	182
6.6 Unidades de observação em várias tabelas.....	185
6.6.1 Carregando vários arquivos usando um laço .....	188
6.6.2 Carregando vários arquivos usando uma list comprehension .....	189
6.7 Conclusão.....	190

<b>Parte III = Manipulação de dados .....</b>	<b>191</b>
<b>Capítulo 7 = Tipos de dados.....</b>	<b>192</b>
7.1 Introdução.....	192
7.2 Tipos de dados .....	192
7.3 Convertendo tipos .....	193
7.3.1 Convertendo para objetos string .....	194
7.3.2 Convertendo para valores numéricos.....	194
7.4 Dados categorizados .....	200
7.4.1 Conversão para categoria.....	201
7.4.2 Manipulando dados categorizados .....	202
7.5 Conclusão.....	203
<b>Capítulo 8 = Strings e dados do tipo texto.....</b>	<b>204</b>
8.1 Introdução .....	204
8.2 Strings .....	205
8.2.1 Obtendo subconjuntos e fatiando strings.....	205
8.2.2 Obtendo o último caractere de uma string .....	207
8.3 Métodos de string .....	209
8.4 Outros métodos de string .....	210
8.4.1 Método join .....	211
8.4.2 Método splitlines .....	211
8.5 Formatação de strings .....	212
8.5.1 Formatação de strings personalizada .....	213
8.5.2 Formatação de strings de caracteres .....	213
8.5.3 Formatação de números .....	214
8.5.4 Formatação no estilo do printf de C .....	215
8.5.5 Strings literais formatadas em Python 3.6+ .....	215
8.6 Expressões regulares (RegEx) .....	216
8.6.1 Correspondência de padrão .....	218
8.6.2 Encontrando um padrão .....	221
8.6.3 Substituindo um padrão.....	221
8.6.4 Compilando um padrão .....	222
8.7 Biblioteca regex .....	224
8.8 Conclusão .....	224
<b>Capítulo 9 = Apply .....</b>	<b>225</b>
9.1 Introdução.....	225
9.2 Funções .....	225
9.3 apply (básico).....	226
9.3.1 apply em uma Series.....	227
9.3.2 apply em um DataFrame.....	229

94 apply (mais avançado) .....	232
94.1 Operações em colunas.....	235
94.2 Operações em linhas .....	237
95 Funções vetorizadas.....	240
95.1 Usando o numpy .....	241
95.2 Usando a biblioteca numba .....	242
96 Funções lambda .....	243
97 Conclusão .....	245

## **Capítulo 10 = Operações groupby: separar–aplicar–combinar ..... 246**

10.1 Introdução .....	246
10.2 Agregação.....	247
10.2.1 Agregação básica com agrupamento de uma única variável .....	247
10.2.2 Métodos de agregação embutidos.....	249
10.2.3 Funções de agregação.....	250
10.2.4 Várias funções simultaneamente.....	254
10.2.5 Usando um dicionário em agg/aggregate .....	254
10.3 Transformação .....	256
10.3.1 Exemplo com escore z.....	256
10.4 Filtragem .....	261
10.5 Objeto pandas.core.groupby.DataFrameGroupBy .....	262
10.5.1 Grupos .....	263
10.5.2 Cálculos com grupos envolvendo diversas variáveis.....	264
10.5.3 Seleccionando um grupo .....	265
10.5.4 Iterando nos grupos.....	265
10.5.5 Vários grupos .....	268
10.5.6 Obtendo resultados planos.....	268
10.6 Trabalhando com MultiIndex .....	269
10.7 Conclusão .....	273

## **Capítulo 11 = Tipo de dado datetime ..... 274**

11.1 Introdução .....	274
11.2 Objeto datetime de Python.....	275
11.3 Conversão para datetime .....	275
11.4 Carregando dados que incluam datas .....	279
11.5 Extraíndo componentes de datas .....	280
11.6 Cálculos com datas e timedeltas .....	282
11.7 Métodos de datetime.....	284
11.8 Obtendo dados de ações.....	287
11.9 Obtendo subconjuntos de dados com base em datas .....	288
11.9.1 Objeto DatetimeIndex.....	289
11.9.2 Objeto TimedeltaIndex .....	290

11.10 Intervalos de datas .....	291
11.10.1 Frequências .....	293
11.10.2 Offsets.....	294
11.11 Deslocando valores .....	295
11.12 Reamostragem .....	303
11.13 Fusos horários .....	304
11.14 Conclusão .....	306
<b>Parte IV = Modelagem de dados.....</b>	<b>307</b>
<b>Capítulo 12 = Modelos lineares.....</b>	<b>308</b>
12.1 Introdução .....	308
12.2 Regressão linear simples.....	308
12.2.1 Usando a statsmodels.....	309
12.2.2 Usando a sklearn.....	311
12.3 Regressão múltipla.....	313
12.3.1 Usando a statsmodels .....	313
12.3.2 Usando a statsmodels com variáveis categorizadas .....	314
12.3.3 Usando a sklearn .....	316
12.3.4 Usando a sklearn com variáveis categorizadas .....	317
12.4 Mantendo os rótulos dos índices com a sklearn .....	318
12.5 Conclusão .....	319
<b>Capítulo 13 = Modelos lineares generalizados.....</b>	<b>320</b>
13.1 Introdução .....	320
13.2 Regressão logística .....	320
13.2.1 Usando a statsmodels.....	322
13.2.2 Usando a sklearn.....	324
13.3 Regressão de Poisson.....	326
13.3.1 Usando a statsmodels .....	326
13.3.2 Regressão binomial negativa para superdispersão .....	328
13.4 Outros modelos lineares generalizados .....	329
13.5 Análise de sobrevivência.....	330
13.5.1 Testando as suposições do modelo de Cox.....	333
13.6 Conclusão .....	334
<b>Capítulo 14 = Diagnóstico de modelos .....</b>	<b>335</b>
14.1 Introdução .....	335
14.2 Resíduos .....	335
14.2.1 Plotagens q-q .....	338
14.3 Comparando vários modelos.....	340
14.3.1 Trabalhando com modelos lineares .....	340



14.3.2 Trabalhando com modelos GLM .....	344
14.4 Validação cruzada k-fold .....	347
14.5 Conclusão .....	351
<b>Capítulo 15 ■ Regularização .....</b>	<b>352</b>
15.1 Introdução .....	352
15.2 Por que regularizar? .....	352
15.3 Regressão LASSO .....	355
15.4 Regressão de ridge.....	357
15.5 Rede elástica.....	359
15.6 Validação cruzada .....	362
15.7 Conclusão .....	365
<b>Capítulo 16 ■ Clustering.....</b>	<b>366</b>
16.1 Introdução .....	366
16.2 k-means .....	366
16.2.1 Redução de dimensões com PCA .....	369
16.3 Clustering hierárquico .....	374
16.3.1 Clustering completo .....	374
16.3.2 Clustering simples .....	375
16.3.3 Clustering com médias .....	375
16.3.4 Clustering com centroide .....	375
16.3.5 Definindo manualmente o limite .....	376
16.4 Conclusão .....	377
<b>Parte V ■ Conclusão .....</b>	<b>378</b>
<b>Capítulo 17 ■ Vida além do Pandas.....</b>	<b>379</b>
17.1 A pilha de processamento (científico).....	379
17.2 Desempenho .....	380
17.2.1 Medindo o tempo de execução de seu código.....	380
17.2.2 Gerando o perfil de seu código .....	382
17.3 Maior e mais rápido .....	382
<b>Capítulo 18 ■ No caminho para ser autodidata.....</b>	<b>383</b>
18.1 É perigoso andar sozinho! .....	383
18.2 Meetups locais .....	383
18.3 Conferências.....	384
18.4 Internet .....	385
18.5 Podcasts .....	385
18.6 Conclusão .....	385

Parte VI = Apêndices.....	386
Apêndice A = Instalação.....	387
Apêndice B = Linha de comandos .....	389
Apêndice C = Templates de projeto.....	391
Apêndice D = Usando Python .....	392
Apêndice E = Diretórios de trabalho.....	395
Apêndice F = Ambientes.....	397
Apêndice G = Instalação de pacotes .....	400
Apêndice H = Importando bibliotecas.....	402
Apêndice I = Listas.....	404
Apêndice J = Tuplas.....	406
Apêndice K = Dicionários.....	407
Apêndice L = Fatiando valores .....	410
Apêndice M = Laços .....	412
Apêndice N = Comprehensions .....	414
Apêndice O = Funções .....	416
Apêndice P = Intervalos e geradores.....	421
Apêndice Q = Atribuição múltipla .....	424
Apêndice R = ndarray do numpy.....	426
Apêndice S = Classes .....	428
Apêndice T = Odo: o modificador de formato.....	430