

Spis treści

Słowo wstępne do wydania drugiego	15
Słowo wstępne do wydania pierwszego	17
Przedmowa	19
Podziękowania	27
O autorze	31
Zmiany w wydaniu drugim	32
Część I Wprowadzenie	33
Rozdział 1 Typ danych DataFrame biblioteki Pandas — podstawy	35
1.1. Wprowadzenie	35
Cele rozdziału	36
1.2. Ładowanie pierwszego zbioru danych	36
1.3. Sprawdzanie kolumn, wierszy i komórek	39
1.3.1. Wybieranie i określanie podzbioru kolumn na podstawie nazwy	39
1.3.2. Określanie podzbioru wierszy	43
1.3.3. Określanie podzbioru wierszy za pomocą numeru wiersza: atrybut .iloc[]	45
1.3.4. Użycie kombinacji	47
1.3.5. Określanie podzbioru wierszy i kolumn	53
1.4. Obliczenia grupowane i agregowane	54
1.4.1. Średnie grupowane	55
1.4.2. Liczebności grupowane	58
1.5. Podstawowy wykres	58
Podsumowanie	59

Rozdział 2	Struktury danych biblioteki Pandas — podstawy	61
	Cele rozdziału	61
2.1.	Tworzenie własnych danych	61
2.1.1.	Tworzenie obiektu Series	62
2.1.2.	Tworzenie obiektu DataFrame	63
2.2.	Obiekty Series	63
2.2.1.	Obiekt Series przypomina typ ndarray	65
2.2.2.	Określanie podzbioru wartości boolowskich: obiekt Series	67
2.2.3.	Operacje są automatycznie wyrównywane i wektoryzowane (rozgłaszanie)	69
2.3.	Obiekt DataFrame	72
2.3.1.	Części obiektu DataFrame	72
2.3.2.	Określanie podzbioru wartości boolowskich: obiekty DataFrame	73
2.3.3.	Operacje są automatycznie wyrównywane i wektoryzowane (rozgłaszanie)	73
2.4.	Wprowadzanie zmian w obiektach Series i DataFrame	75
2.4.1.	Dodawanie dodatkowych kolumn	75
2.4.2.	Bezpośrednie modyfikowanie kolumny	76
2.4.3.	Modyfikowanie kolumn za pomocą metody .assign()	79
2.4.4.	Usuwanie wartości	81
2.5.	Eksportowanie i importowanie danych	81
2.5.1.	„Peklowanie”	82
2.5.2.	Format danych CSV	84
2.5.3.	Excel	84
2.5.4.	Format Feather	86
2.5.5.	Projekt Arrow	87
2.5.6.	Słownik	87
2.5.7.	Format JSON	88
2.5.8.	Inne typy danych wyjściowych	91
	Podsumowanie	92
Rozdział 3	Tworzenie wykresów — podstawy	93
	Cele rozdziału	93
3.1.	Dlaczego warto wizualizować dane?	93
3.2.	Podstawy obsługi biblioteki matplotlib	94
3.2.1.	Obiekty rysunków i podwykresy z osiami	95
3.2.2.	Anatomia rysunku	99
3.3.	Tworzenie graficznych wizualizacji danych statystycznych za pomocą biblioteki matplotlib	100
3.3.1.	Jednozmienność (pojedyncza zmienna)	101
3.3.2.	Dwuzmienność (dwie zmienne)	101
3.3.3.	Dane wielozmienne	103

3.4. Biblioteka seaborn	105
3.4.1. Jednozmienność	106
3.4.2. Dane dwuzmienne	110
3.4.3. Dane wielozmienne	120
3.4.4. Aspekty	124
3.4.5. Style i kompozycje biblioteki seaborn	129
3.4.6. Jak korzystać z dokumentacji biblioteki seaborn?	133
3.4.7. Interfejs biblioteki seaborn następnej generacji	135
3.5. Metoda tworzenia wykresów za pomocą biblioteki Pandas	136
3.5.1. Histogram	136
3.5.2. Wykres gęstości	137
3.5.3. Wykres punktowy	137
3.5.4. Wykres przedziałów sześciokątnych (hexbin)	138
3.5.5. Wykres pudełkowy	139
Podsumowanie	140
Rozdział 4 Dane uporządkowane	141
Cele rozdziału	142
Uwaga dotycząca niniejszego rozdziału	142
4.1. Kolumny zawierają wartości, a nie zmienne	142
4.1.1. Utrwalenie jednej kolumny	142
4.1.2. Utrwalenie wielu kolumn	145
4.2. Kolumny zawierają wiele zmiennych	146
4.2.1. Osobne dzielenie i dodawanie kolumn	147
4.2.2. Dzielenie i łączenie kolumn w jednym kroku	149
4.3. Zmienne znajdują się w wierszach i kolumnach	150
Podsumowanie	153
Rozdział 5 Zastosowanie funkcji	154
Cele rozdziału	154
Uwaga dotycząca niniejszego rozdziału	155
5.1. Elementarz funkcji	155
5.2. Zastosowanie funkcji (podstawy)	156
5.2.1. Zastosowanie funkcji względem obiektu Series	157
5.2.2. Zastosowanie funkcji względem obiektu DataFrame	158
5.3. Funkcje wektoryzowane	161
5.3.1. Wektoryzacja za pomocą biblioteki NumPy	162
5.3.2. Wektoryzacja za pomocą biblioteki Numba	163
5.4. Funkcje lambda (funkcje anonimowe)	164
Podsumowanie	165

Część II	Przetwarzanie danych	167
Rozdział 6	Łączenie danych	169
	Cele rozdziału	169
	6.1. Łączenie zbiorów danych	169
	6.2. Konkatenacja	170
	6.2.1. Części przeglądowe obiektu DataFrame	171
	6.2.2. Dodawanie wierszy	171
	6.2.3. Dodawanie kolumn	174
	6.2.4. Konkatenacja z różnymi indeksami	175
	6.3. Jednostki obserwacyjne w obrębie wielu tabel	178
	6.3.1. Ładowanie wielu plików za pomocą pętli	180
	6.3.2. Ładowanie wielu plików przy użyciu listy składanej	182
	6.4. Scalanie wielu zbiorów danych	183
	6.4.1. Scalanie typu „jedna z jedną”	185
	6.4.2. Scalanie typu „wiele z jedną”	186
	6.4.3. Scalanie typu „wiele z wieloma”	187
	6.4.4. Sprawdzanie wyników pracy za pomocą asercji	189
	Podsumowanie	189
Rozdział 7	Normalizacja danych	190
	Cele rozdziału	190
	7.1. Wiele jednostek obserwacyjnych w tabeli (normalizacja)	190
	Podsumowanie	194
Rozdział 8	Operacje grupowania: dzielenie, stosowanie i łączenie	195
	Cele rozdziału	196
	8.1. Agregacja	196
	8.1.1. Podstawowa agregacja danych grupowanych z jedną zmienną	196
	8.1.2. Wbudowane metody agregacji	198
	8.1.3. Funkcje agregacji	199
	8.1.4. Użycie wielu funkcji jednocześnie	202
	8.1.5. Zastosowanie słownika w metodzie .agg() lub .aggregate()	202
	8.2. Transformacja	204
	8.2.1. Przykład wyniku standardowego z	204
	8.2.2. Przykład z brakującymi wartościami	206
	8.3. Filtrowanie	208
	8.4. Obiekt pandas.core.groupby.DataFrameGroupBy	209
	8.4.1. Grupy	209
	8.4.2. Obliczenia w ramach grupowania obejmujące wiele zmiennych	210
	8.4.3. Wybieranie grupy	211
	8.4.4. Iteracja w obrębie grup	211

8.4.5. Wiele grup	213
8.4.6. „Spłaszczanie” wyników (.reset_index())	213
8.5. Zastosowanie obiektu MultiIndex	214
Podsumowanie	218

Część III Typy danych219

Rozdział 9 Brakujące dane 221

Cele rozdziału	221
9.1. Czym jest wartość NaN?	221
9.2. Skąd biorą się brakujące wartości?	223
9.2.1. Ładowanie danych	223
9.2.2. Scalone dane	224
9.2.3. Wartości wprowadzane przez użytkownika	225
9.2.4. Ponowne indeksowanie	226
9.3. Zajmowanie się brakującymi danymi	227
9.3.1. Znajdowanie brakujących danych i określanie ich ilości	228
9.3.2. Oczyszczanie danych z brakującymi wartościami	229
9.3.3. Obliczenia uwzględniające brakujące dane	233
9.4. Brakująca wartość NA wbudowana w bibliotecz Pandas	234
Podsumowanie	235

Rozdział 10 Typy danych 236

Cele rozdziału	236
10.1. Typy danych	236
10.2. Przekształcanie typów	237
10.2.1. Konwersja do postaci obiektów łańcuchów	237
10.2.2. Przekształcanie w wartości liczbowe	238
10.3. Dane kategoryjne	242
10.3.1. Przekształcanie w kategorię	242
10.3.2. Przetwarzanie danych kategoryjnych	243
Podsumowanie	244

Rozdział 11 Łańcuchy i dane tekstowe 245

Wprowadzenie	245
Cele rozdziału	245
11.1. Łańcuchy	245
11.1.1. Określanie podzbioru i dzielenie łańcuchów	246
11.1.2. Uzyskanie ostatniego znaku łańcucha	247
11.2. Metody łańcuchowe	249
11.3. Dodatkowe metody łańcuchowe	251
11.3.1. Metoda join	251
11.3.2. Metoda splitlines	251

11.4. Formatowanie łańcuchów (f-łańcuchy)	253
11.4.1. Formatowanie liczb	254
11.5. Wyrażenia regularne	255
11.5.1. Dopasowanie wzorca	257
11.5.2. Pamiętaj, jakich używasz wzorców wyrażeń regularnych	259
11.5.3. Znajdowanie wzorca	261
11.5.4. Zastępowanie wzorca	261
11.5.5. Kompilowanie wzorca	262
11.6. Biblioteka regex	263
Podsumowanie	264
Rozdział 12 Daty i godziny	265
Cele rozdziału	265
12.1. Obiekt datetime języka Python	265
12.2. Przekształcanie do postaci ramki danych	266
12.3. Ładowanie danych zawierających daty	269
12.4. Wyodrębnianie składników daty	270
12.5. Obliczenia obejmujące daty i obiekty timedelta	273
12.6. Metody obiektu datetime	274
12.7. Uzyskiwanie danych notowań giełdowych	277
12.8. Określanie podzbioru danych na podstawie dat	278
12.8.1. Obiekt DatetimeIndex	279
12.8.2. Obiekt TimedeltaIndex	280
12.9. Zakresy dat	281
12.9.1. Częstotliwości	283
12.9.2. Przesunięcia	284
12.10. Wartości przesuwające	284
12.11. Ponowne próbkowanie	290
12.12. Strefy czasowe	292
12.13. Biblioteka Arrow do lepszej obsługi dat i godzin	294
Podsumowanie	294
Część IV Modelowanie danych	295
Rozdział 13 Regresja liniowa (wynikowa zmienna ciągła)	297
13.1. Prosta regresja liniowa	297
13.1.1. Użycie biblioteki statsmodels	298
13.1.2. Zastosowanie biblioteki scikit-learn (sklearn)	299
13.2. Regresja wielokrotna	301
13.2.1. Użycie biblioteki statsmodels	301
13.2.2. Zastosowanie biblioteki scikit-learn (sklearn)	302

13.3. Modele ze zmiennymi kategorialnymi	303
13.3.1. Zmienne kategorialne w bibliotece statsmodels	303
13.3.2. Zmienne kategorialne w bibliotece scikit-learn (sklearn)	305
13.4. Kodowanie One-Hot w bibliotece scikit-learn z wykorzystaniem potoków transformera	307
Podsumowanie	309
Rozdział 14 Uogólnione modele liniowe	310
Coś o tym rozdziale	310
14.1. Regresja logistyczna (binarna zmienna wyjściowa)	310
14.1.1. Użycie biblioteki statsmodels	312
14.1.2. Zastosowanie biblioteki sklearn	314
14.1.3. Zachowaj ostrożność w przypadku domyślnych wartości biblioteki scikit-learn (sklearn)	315
14.2. Regresja Poissona (ilościowa zmienna wynikowa)	317
14.2.1. Użycie biblioteki statsmodels	317
14.2.2. Ujemna regresja dwumianowa w przypadku nadmiernej dyspersji	319
14.3. Bardziej uogólnione modele liniowe	321
Podsumowanie	322
Rozdział 15 Analiza przeżycia	323
15.1. Dane analizy przeżycia	324
15.2. Krzywe Kaplana-Meiera	324
15.3. Model proporcjonalnego hazardu Coxa	326
15.3.1. Testowanie założeń modelu Coxa	327
Podsumowanie	328
Rozdział 16 Diagnostyka modeli	329
16.1. Residua	329
16.1.1. Wykresy kwantylowe K-K	332
16.2. Porównanie wielu modeli	334
16.2.1. Korzystanie z modeli liniowych	334
16.2.2. Zastosowanie uogólnionych modeli liniowych	337
16.3. Walidacja krzyżowa k-krotna	339
Podsumowanie	342
Rozdział 17 Regularyzacja	343
17.1. Dlaczego regularyzacja?	343
17.2. Regresja LASSO	345
17.3. Regresja grzbietowa	346
17.4. Sieć elastyczna	347
17.5. Walidacja krzyżowa	349
Podsumowanie	350

Rozdział 18	Klasteryzacja	351
18.1.	k-średnie	351
18.1.1.	Ograniczanie liczby wymiarów za pomocą analizy PCA	353
18.2.	Klastrowanie hierarchiczne	357
18.2.1.	Klastrowanie kompletne	358
18.2.2.	Klastrowanie pojedyncze	358
18.2.3.	Klastrowanie ze średnią	359
18.2.4.	Klastrowanie z centroidem	360
18.2.5.	Klastrowanie metodą Warda	360
18.2.6.	Ręczne ustawianie progu	361
	Podsumowanie	361
Część V	Podsumowanie	363
Rozdział 19	Świat poza obrębem biblioteki Pandas	365
19.1.	Stos do obliczeń (naukowych)	365
19.2.	Wydajność	366
19.2.1.	Pomiar czasu wykonywania kodu	366
19.2.2.	Profilowanie kodu	366
19.2.3.	Moduł concurrent.futures	366
19.3.	Dask	367
19.4.	Siuba	367
19.5.	Ibis	367
19.6.	Polars	367
19.7.	PyJanitor	368
19.8.	Pandera	368
19.9.	Uczenie maszynowe	368
19.10.	Publikowanie	368
19.11.	Panele kontrolne	369
	Podsumowanie	369
Rozdział 20	Działanie w pojedynkę jest niebezpieczne!	370
20.1.	Lokalne spotkania	370
20.2.	Konferencje	371
20.3.	The Carpentries	371
20.4.	Podcasty	372
20.5.	Inne zasoby	372
	Podsumowanie	372

Dodatki	373
Dodatek A	Mapy pojęć
Dodatek B	Instalacja i konfiguracja
B.1.	Instalacja języka Python
B.1.1.	Anaconda
B.1.2.	Miniconda
B.1.3.	Odinstalowywanie dystrybucji Anaconda lub Miniconda
B.1.4.	pyenv
B.2.	Instalowanie pakietów języka Python
B.3.	Pobieranie zbiorów danych używanych w książce
Dodatek C	Wiersz poleceń
C.1.	Instalacja
C.1.1.	System Windows
C.1.2.	System Mac
C.1.3.	System Linux
C.2.	Podstawy
Dodatek D	Szablony projektowe
Dodatek E	Zastosowanie języka Python
E.1.	Wiersz poleceń i edytor tekstu
E.2.	Python i IPython
E.3.	Jupyter
E.4.	Zintegrowane środowiska programistyczne IDE
Dodatek F	Katalogi robocze
Dodatek G	Środowiska
G.1.	Środowiska systemu conda
G.2.	Pyenv + Pipenv
Dodatek H	Instalacja pakietów
H.1.	Aktualizowanie pakietów
Dodatek I	Importowanie bibliotek
Dodatek J	Styl kodu
J.1.	Znaki podziału wiersza w kodzie
Dodatek K	Kontenery: listy, krotki i słowniki
K.1.	Listy
K.2.	Krotki
K.3.	Słowniki

Dodatek L	Określanie wartości za pomocą składni wycinków	404
Dodatek M	Pętle	406
Dodatek N	Listy składane	408
Dodatek O	Funkcje	410
	O.1. Parametry domyślne	412
	O.2. Parametry arbitralne	412
	O.2.1. Wyrażenie *args	413
	O.2.2. Wyrażenie **kwargs	413
Dodatek P	Zakresy i generatory	414
Dodatek Q	Przypisanie wielokrotne	416
Dodatek R	Typ ndarray biblioteki NumPy	418
Dodatek S	Klasy	420
Dodatek T	Komunikat SettingWithCopyWarning	422
	T.1. Modyfikowanie podzbioru danych	422
	T.2. Zastępowanie wartości	424
	T.3. Dodatkowe zasoby informacji	425
Dodatek U	Tworzenie łańcuchów metod	426
Dodatek V	Czas wykonywania kodu	428
Dodatek W	Formatowanie łańcuchów	430
	W.1. Formatowanie w stylu języka C	430
	W.2. Formatowanie łańcuchów: metoda .format()	431
	W.3. Formatowanie liczb	431
Dodatek X	Instrukcje warunkowe (if-elif-else)	433
Dodatek Y	Przykład regresji logistycznej ze zbiorem danych ACS dla Nowego Jorku	435
	Y.0.1. Użycie biblioteki sklearn	439
Dodatek Z	Replikowanie wyników za pomocą języka R	442
	Z.1. Regresja liniowa	443
	Z.2. Regresja logistyczna	445
	Z.3. Regresja Poissona	446
	Z.3.1. Ujemna regresja dwumianowa w przypadku nadmiernej dyspersji	446
Skorowidz		449