

Spis treści

Przedmowa	11
1. Wstęp	17
1.1. O czym jest ta książka?	17
Jakie rodzaje danych?	17
1.2. Dlaczego warto korzystać z Pythona w celu przeprowadzenia analizy danych?	18
Python jako spoiwo	18
Rozwiązywanie problemu „dwujęzyczności”	19
Dlaczego nie Python?	19
1.3. Podstawowe biblioteki Pythona	20
NumPy	20
pandas	21
Matplotlib	22
IPython i Jupyter	22
SciPy	23
Scikit-learn	23
statsmodels	24
Inne pakiety	25
1.4. Instalacja i konfiguracja	25
Windows	25
GNU, Linux	26
macOS	26
Instalacja niezbędnych pakietów	27
Zintegrowane środowiska programistyczne i edytory tekstowe	28
1.5. Społeczność i konferencje	28
1.6. Nawigacja po książce	29
Przykłady kodu	30
Przykładowe dane	30
Konwencje importowania	31

2. Podstawy Pythona oraz obsługi narzędzi IPython i Jupyter	32
2.1. Interpreter Pythona	33
2.2. Podstawy interpretera IPython	34
Uruchamianie powłoki IPython	34
Uruchamianie notatnika Jupyter Notebook	35
Uzupełnianie poleceń	38
Introspekcja	39
2.3. Podstawy Pythona	40
Semantyka języka Python	40
Skalarne typy danych	48
Przepływ sterowania	55
2.4. Podsumowanie	58
3. Wbudowane struktury danych, funkcje i pliki	59
3.1. Struktury danych i sekwencje	59
Krotka	59
Lista	62
Słownik	66
Zbiór	70
Wbudowane funkcje obsługujące sekwencje	72
Lista, słownik i zbiór — składanie	74
3.2. Funkcje	76
Przestrzenie nazw, zakres i funkcje lokalne	77
Zwracanie wielu wartości	78
Funkcje są obiektami	79
Funkcje anonimowe (lambda)	80
Generatory	81
Błędy i obsługa wyjątków	83
3.3. Pliki i system operacyjny	86
Bajty i kodowanie Unicode w plikach	89
3.4. Podsumowanie	91
4. Podstawy biblioteki NumPy: obsługa tablic i wektorów	92
4.1. NumPy ndarray — wielowymiarowy obiekt tablicowy	94
Tworzenie tablic ndarray	95
Typ danych tablic ndarray	97
Działania matematyczne z tablicami NumPy	100
Podstawy indeksowania i przechwytywania części	101
Indeksowanie i wartości logiczne	106
Indeksowanie specjalne	108
Transponowanie tablic i zamiana osi	110

4.2.	Generowanie liczb pseudolosowych	111
4.3.	Funkcje uniwersalne — szybkie funkcje wykonywane na poszczególnych elementach tablicy	113
4.4.	Programowanie z użyciem tablic	115
	Logiczne operacje warunkowe jako operacje tablicowe	117
	Metody matematyczne i statystyczne	119
	Metody tablic logicznych	120
	Sortowanie	121
	Wartości unikalne i operacje logiczne	122
4.5.	Tablice i operacje na plikach	123
4.6.	Algebra liniowa	124
4.7.	Przykład: błędzenie losowe	125
	Jednoczesne symulowanie wielu błędzeń losowych	127
4.8.	Podsumowanie	128
5.	Rozpoczynamy pracę z biblioteką pandas	129
5.1.	Wprowadzenie do struktur danych biblioteki pandas	130
	Obiekt Series	130
	Obiekt DataFrame	134
	Obiekty index	141
5.2.	Podstawowe funkcjonalności	142
	Uaktualnianie indeksu	143
	Odrzucanie elementów osi	145
	Indeksowanie, wybieranie i filtrowanie	147
	Działania arytmetyczne i wyrównywanie danych	156
	Funkcje apply i map	161
	Sortowanie i tworzenie rankingów	163
	Indeksy osi ze zduplikowanymi etykietami	167
5.3.	Podsumowywanie i generowanie statystyk opisowych	168
	Współczynnik korelacji i kowariancja	171
	Unikalne wartości, ich liczba i przynależność	173
5.4.	Podsumowanie	176
6.	Odczyt i zapis danych, formaty plików	177
6.1.	Odczyt i zapis danych w formacie tekstowym	177
	Wczytywanie części pliku tekstowego	184
	Zapis danych w formacie tekstowym	185
	Praca z plikami danych rozgraniczonych	187
	Dane w formacie JSON	189
	XML i HTML — web scraping	190

6.2.	Formaty danych binarnych	194
	Wczytywanie plików programu Microsoft Excel	195
	Obsługa formatu HDF5	196
6.3.	Obsługa interfejsów sieciowych	198
6.4.	Obsługa baz danych	200
6.5.	Podsumowanie	202
7.	Czyszczenie i przygotowywanie danych	203
7.1.	Obsługa brakujących danych	203
	Filtrowanie brakujących danych	205
	Wypełnianie brakujących danych	207
7.2.	Przekształcanie danych	209
	Usuwanie duplikatów	209
	Przekształcanie danych przy użyciu funkcji lub mapowania	210
	Zastępowanie wartości	212
	Zmiana nazw indeksów osi	213
	Dyskretyzacja i podział na koszyki	214
	Wykrywanie i filtrowanie elementów odstających	216
	Permutacje i próbkowanie losowe	218
	Przetwarzanie wskaźników i zmiennych zastępczych	219
7.3.	Rozszerzone typy danych	222
7.4.	Operacje przeprowadzane na łańcuchach	225
	Metody obiektu typu string	225
	Wyrażenia regularne	227
	Funkcje tekstowe w pakiecie pandas	230
7.5.	Dane kategoryczne	233
	Kontekst i motywacja	233
	Rozszerzony typ Categorical w bibliotece pandas	235
	Obliczenia na obiektach typu Categorical	237
	Metody obiektu kategorycznego	239
7.6.	Podsumowanie	242
8.	Przetwarzanie danych — operacje łączenia, wiązania i przekształcania	243
8.1.	Indeksowanie hierarchiczne	243
	Zmiana kolejności i sortowanie poziomów	246
	Parametry statystyki opisowej z uwzględnieniem poziomu	247
	Indeksowanie z kolumnami ramki danych	247
8.2.	Łączenie zbiorów danych	249
	Łączenie ramek danych w stylu łączenia elementów baz danych	249
	Łączenie przy użyciu indeksu	253
	Konkatenacja wzdłuż osi	258
	Łączenie częściowo nakładających się danych	263

8.3.	Zmiana kształtu i operacje osiowe	264
	Przekształcenia z indeksowaniem hierarchicznym	264
	Przekształcanie z formatu „długiego” na „szeroki”	267
	Przekształcanie z formatu „szerokiego” na „długi”	270
8.4.	Podsumowanie	272
9.	Wykresy i wizualizacja danych	273
9.1.	Podstawy obsługi interfejsu pakietu matplotlib	274
	Obiekty figure i wykresy składowe	275
	Kolory, oznaczenia i style linii	279
	Punkty, etykiety i legendy	281
	Adnotacje i rysunki	283
	Zapisywanie wykresów w postaci plików	286
	Konfiguracja pakietu matplotlib	287
9.2.	Generowanie wykresów za pomocą pakietów pandas i seaborn	287
	Wykresy liniowe	287
	Wykresy słupkowe	290
	Histogramy i wykresy gęstości	296
	Wykresy punktowe	298
	Wykresy panelowe i dane kategoryczne	300
9.3.	Inne narzędzia przeznaczone do wizualizacji danych w Pythonie	301
9.4.	Podsumowanie	303
10.	Agregacja danych i operacje wykonywane na grupach	304
10.1.	Mechanika interfejsu groupby	305
	Iteracja po grupach	308
	Wybieranie kolumny lub podzbioru kolumn	310
	Grupowanie przy użyciu słowników i serii	311
	Grupowanie przy użyciu funkcji	312
	Grupowanie przy użyciu poziomów indeksu	312
10.2.	Agregacja danych	313
	Przetwarzanie kolumna po kolumnie i stosowanie wielu funkcji	315
	Zwracanie zagregowanych danych bez indeksów wierszy	318
10.3.	Metoda apply — ogólne zastosowanie techniki dziel-zastosuj-połącz	319
	Usuwanie kluczy grup	321
	Kwantyle i analiza koszykowa	321
	Przykład: wypełnianie brakujących wartości przy użyciu wartości charakterystycznych dla grupy	323
	Przykład: losowe generowanie próbek i permutacja	325
	Przykład: średnie ważone grup i współczynnik korelacji	327
	Przykład: regresja liniowa grup	329
10.4.	Transformacje grup i „nieobudowane” operacje grupowania	330

10.5. Tabele przestawne i krzyżowe	333
Tabele krzyżowe	336
10.6. Podsumowanie	337
11. Szeregi czasowe	338
11.1. Typy danych i narzędzia przeznaczone do obsługi daty i czasu	339
Konwersja pomiędzy obiektami string i datetime	340
11.2. Podstawy szeregów czasowych	342
Indeksowanie i wybieranie	343
Szeregi czasowe z duplikatami indeksów	346
11.3. Zakresy dat, częstotliwości i przesunięcia	347
Generowanie zakresów dat	347
Częstotliwości i przesunięcia daty	350
Przesuwanie daty	351
11.4. Obsługa strefy czasowej	354
Lokalizacja i konwersja stref czasowych	355
Operacje z udziałem obiektów Timestamp o wyznaczonej strefie czasowej	357
Operacje pomiędzy różnymi strefami czasowymi	358
11.5. Okresy i przeprowadzanie na nich operacji matematycznych	358
Konwersja częstotliwości łańcuchów	359
Kwartalne częstotliwości okresów	361
Konwersja znaczników czasu na okresy (i z powrotem)	362
Tworzenie obiektów PeriodIndex na podstawie tablic	364
11.6. Zmiana rozdzielczości i konwersja częstotliwości	365
Zmniejszanie częstotliwości	366
Zwiększanie rozdzielczości i interpolacja	369
Zmiana rozdzielczości z okresami	371
Grupowa zmiana częstotliwości	372
11.7. Funkcje ruchomego okna	374
Funkcje ważone wykładniczo	377
Binarne funkcje ruchomego okna	378
Funkcje ruchomego okna definiowane przez użytkownika	380
11.8. Podsumowanie	381
12. Wprowadzenie do bibliotek modelujących	382
12.1. Łączenie pandas z kodem modelu	382
12.2. Tworzenie opisów modeli przy użyciu biblioteki Patsy	385
Przekształcania danych za pomocą formuł Patsy	387
Patsy i dane katagoryczne	389
12.3. Wprowadzenie do biblioteki statsmodels	391
Szacowanie modeli liniowych	392
Szacowanie procesów szeregów czasowych	394

12.4.	Wprowadzenie do pakietu scikit-learn	395
12.5.	Podsumowanie	399
13.	Przykłady analizy danych	400
13.1.	Dane USA.gov serwisu Bitly	400
	Liczenie stref czasowych w czystym Pythonie	401
	Liczenie stref czasowych przy użyciu pakietu pandas	403
13.2.	Zbiór danych MovieLens 1M	409
	Wyznaczenie rozbieżności ocen	413
13.3.	Imiona nadawane dzieciom w USA w latach 1880 – 2010	416
	Analiza trendów imion	421
13.4.	Baza danych USDA Food	430
13.5.	Baza danych 2012 Federal Election Commission	436
	Statystyki datków z podziałem na wykonywany zawód i pracodawcę	438
	Podział kwot datków na koszyki	441
	Statystyki datków z podziałem na poszczególne stany	443
13.6.	Podsumowanie	444
A.	Zaawansowane zagadnienia związane z biblioteką NumPy	445
A.1.	Szczegóły budowy obiektu ndarray	445
	Hierarchia typów danych NumPy	446
A.2.	Zaawansowane operacje tablicowe	447
	Zmiana wymiarów tablic	448
	Kolejności charakterystyczne dla języków C i Fortran	449
	Łączenie i dzielenie tablic	450
	Powtarzanie elementów — funkcje tile i repeat	452
	Alternatywy indeksowania specjalnego — metody take i put	454
A.3.	Rozgłaszanie	455
	Rozgłaszanie wzdłuż innych osi	457
	Przypisywanie wartości elementom tablicy poprzez rozgłaszanie	460
A.4.	Zaawansowane zastosowania funkcji uniwersalnych	461
	Metody instancji funkcji uniwersalnych	461
	Pisanie nowych funkcji uniwersalnych w Pythonie	463
A.5.	Tablice o złożonej strukturze	464
	Zagnieżdżone typy danych i pola wielowymiarowe	464
	Do czego przydają się tablice o złożonej strukturze?	465
A.6.	Jeszcze coś o sortowaniu	466
	Sortowanie pośrednie — metody argsort i lexsort	467
	Alternatywne algorytmy sortowania	468
	Częściowe sortowanie tablic	469
	Wyszukiwanie elementów w posortowanej tablicy za pomocą metody numpy.searchsorted	470

A.7.	Pisanie szybkich funkcji NumPy za pomocą pakietu Numba	471
	Tworzenie obiektów <code>numpy.ufunc</code> za pomocą pakietu Numba	472
A.8.	Zaawansowane tablicowe operacje wejścia i wyjścia	473
	Pliki mapowane w pamięci	473
	HDF5 i inne możliwości zapisu tablic	474
A.9.	Jak zachować wysoką wydajność?	475
	Dlaczego warto korzystać z sąsiadujących ze sobą obszarów pamięci?	475
B.	Dodatkowe informacje dotyczące systemu IPython	477
B.1.	Skróty klawiaturowe	477
B.2.	Magiczne polecenia	478
	Polecenie <code>%run</code>	480
	Uruchamianie kodu zapisanego w schowku	481
B.3.	Korzystanie z historii poleceń	482
	Przeszukiwanie i korzystanie z historii poleceń	482
	Zmienne wejściowe i wyjściowe	483
B.4.	Interakcja z systemem operacyjnym	484
	Polecenia powłoki systemowej i aliasy	484
	System tworzenia skrótów do katalogów	485
B.5.	Narzędzia programistyczne	486
	Interaktywny debugger	486
	Pomiar czasu — funkcje <code>%time</code> i <code>%timeit</code>	490
	Podstawowe profilowanie — funkcje <code>%prun</code> i <code>%run-p</code>	492
	Profilowanie funkcji linia po linii	494
B.6.	Wskazówki dotyczące produktywnego tworzenia kodu w środowisku IPython	496
	Przeladowywanie modułów	496
	Wskazówki dotyczące projektowania kodu	497
B.7.	Zaawansowane funkcje środowiska IPython	498
	Profile i konfiguracja	498
B.8.	Podsumowanie	500