

**Python w analizie danych : przetwarzanie danych za pomocą pakietów
Pandas i NumPy oraz środowiska IPython / Wes McKinney. – Gliwice, ©
2018**

Spis treści

Przedmowa	11
1. Wstęp	15
1.1. O czym jest ta książka?	15
Jakie rodzaje danych?	15
1.2. Dlaczego warto korzystać z Pythona w celu przeprowadzenia analizy danych?	16
Python jako spoiwo	16
Rozwiązywanie problemu „dwujęzyczności”	17
Dlaczego nie Python?	17
1.3. Podstawowe biblioteki Pythona	17
NumPy	18
pandas	18
Matplotlib	19
IPython i Jupyter	19
SciPy	20
Scikit-learn	21
statsmodels	21
1.4. Instalacja i konfiguracja	22
Windows	22
Apple (OS X, macOS)	23
GNU, Linux	23
Instalowanie i aktualizowanie pakietów Pythona	24
Python 2 i Python 3	24
Zintegrowane środowiska programistyczne i edytory tekstowe	25
1.5. Społeczność i konferencje	25
1.6. Nawigacja po książce	26
Przykłady kodu	27
Przykładowe dane	27
Konwencje importowania	27
Żargon	27
2. Podstawy Pythona oraz obsługi narzędzi IPython i Jupyter	29
2.1. Interpreter Pythona	30
2.2. Podstawy interpretera IPython	31
Uruchamianie powłoki IPython	31
Uruchamianie notatnika Jupyter Notebook	32
Uzupełnianie poleceń	35

Introspekcja	36
Polecenie %run	37
Wykonywanie kodu ze schowka	39
Skróty klawiaturowe działające w terminalu	39
Polecenia magiczne	40
Integracja pakietu matplotlib	42
2.3. Podstawy Pythona	42
Semantyka języka Python	43
Skalarne typy danych	50
Przepływ sterowania	57
3. Wbudowane struktury danych, funkcje i pliki	61
3.1. Struktury danych i sekwencje	61
Krotka	61
Lista	64
Wbudowane funkcje obsługujące sekwencje	68
Słownik	70
Zbiór	73
Lista, słownik i zbiór — składanie	75
3.2. Funkcje	77
Przestrzenie nazw, zakres i funkcje lokalne	78
Zwracanie wielu wartości	79
Funkcje są obiektami	79
Funkcje anonimowe (lambda)	81
Currying — częściowa aplikacja argumentów	82
Generatory	82
Błędy i obsługa wyjątków	84
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	93
4.1. NumPy ndarray — wielowymiarowy obiekt tablicowy	95
Tworzenie tablic ndarray	96
Typ danych tablic ndarray	98
Działania matematyczne z tablicami NumPy	100
Podstawy indeksowania i przechwytywania części	101
Indeksowanie i wartości logiczne	105
Indeksowanie specjalne	108
Transponowanie tablic i zamiana osi	109
4.2. Funkcje uniwersalne — szybkie funkcje wykonywane na poszczególnych elementach tablicy	110
4.3. Programowanie z użyciem tablic	113
Logiczne operacje warunkowe jako operacje tablicowe	115
Metody matematyczne i statystyczne	116
Metody tablic logicznych	117

Sortowanie	118
Wartości unikalne i operacje logiczne	119
4.4. Tablice i operacje na plikach	120
4.5. Algebra liniowa	120
4.6. Generowanie liczb pseudolosowych	122
4.7. Przykład: błędzenie losowe	124
Jednoczesne symulowanie wielu błędzeń losowych	125
4.8. Podsumowanie	126
5. Rozpoczynamy pracę z biblioteką pandas	127
5.1. Wprowadzenie do struktur danych biblioteki pandas	127
Obiekt Series	128
Obiekt DataFrame	131
Obiekty index	137
5.2. Podstawowe funkcjonalności	139
Uaktualnianie indeksu	139
Odrzucanie elementów osi	141
Indeksowanie, wybieranie i filtrowanie	143
Indeksy w postaci liczb całkowitych	147
Działania arytmetyczne i wyrównywanie danych	148
Funkcje apply i map	153
Sortowanie i tworzenie rankingów	154
Indeksy osi ze zduplikowanymi etykietami	157
5.3. Podsumowywanie i generowanie statystyk opisowych	158
Współczynnik korelacji i kowariancja	161
Unikalne wartości, ich liczba i przynależność	163
5.4. Podsumowanie	165
6. Odczyt i zapis danych, formaty plików	167
6.1. Odczyt i zapis danych w formacie tekstowym	167
Wczytywanie części pliku tekstowego	173
Zapis danych w formacie tekstowym	174
Praca z plikami danych rozgraniczonych	176
Dane w formacie JSON	178
XML i HTML — web scraping	179
6.2. Formaty danych binarnych	182
Obsługa formatu HDF5	183
Wczytywanie plików programu Microsoft Excel	185
6.3. Obsługa interfejsów sieciowych	186
6.4. Obsługa baz danych	187
6.5. Podsumowanie	188
7. Czyszczenie i przygotowywanie danych	189
7.1. Obsługa brakujących danych	189
Filtrowanie brakujących danych	191
Wypełnianie brakujących danych	193

7.2. Przekształcanie danych	195
Usuwanie duplikatów	195
Przekształcanie danych przy użyciu funkcji lub mapowania	196
Zastępowanie wartości	197
Zmiana nazw indeksów osi	199
Dyskretyzacja i podział na koszyki	200
Wykrywanie i filtrowanie elementów odstających	202
Permutacje i próbkowanie losowe	203
Przetwarzanie wskaźników i zmiennych zastępczych	204
7.3. Operacje przeprowadzane na łańcuchach	207
Metody obiektu typu string	207
Wyrażenia regularne	209
Wektoryzacja funkcji łańcuchów w pakiecie pandas	212
7.4. Podsumowanie	215

8. Przetwarzanie danych — operacje łączenia, wiązania i przekształcania **217**

8.1. Indeksowanie hierarchiczne	217
Zmiana kolejności i sortowanie poziomów	220
Parametry statystyki opisowej z uwzględnieniem poziomu	220
Indeksowanie z kolumnami ramki danych	221
8.2. Łączenie zbiorów danych	222
Łączenie ramek danych w stylu łączenia elementów baz danych	222
Łączenie przy użyciu indeksu	227
Konkatenacja wzdłuż osi	230
Łączenie częściowo nakładających się danych	234
8.3. Zmiana kształtu i operacje osiowe	235
Przekształcenia z indeksowaniem hierarchicznym	236
Przekształcanie z formatu „długiego” na „szeroki”	238
Przekształcanie z formatu „szerokiego” na „długi”	241
8.4. Podsumowanie	243

9. Wykresy i wizualizacja danych **245**

9.1. Podstawy obsługi interfejsu pakietu matplotlib	245
Obiekty figure i wykresy składowe	246
Kolory, oznaczenia i style linii	250
Punkty, etykiety i legendy	252
Adnotacje i rysunki	255
Zapisywanie wykresów w postaci plików	257
Konfiguracja pakietu matplotlib	258
9.2. Generowanie wykresów za pomocą pakietów pandas i seaborn	259
Wykresy liniowe	259
Wykresy słupkowe	262
Histogramy i wykresy gęstości	266
Wykresy punktowe	268
Wykresy panelowe i dane kategoryczne	269

9.3. Inne narzędzia przeznaczone do wizualizacji danych w Pythonie	272
9.4. Podsumowanie	272
10. Agregacja danych i operacje wykonywane na grupach	273
10.1. Mechanika interfejsu groupby	274
Iteracja po grupach	277
Wybieranie kolumny lub podzbioru kolumn	278
Grupowanie przy użyciu słowników i serii	279
Grupowanie przy użyciu funkcji	280
Grupowanie przy użyciu poziomów indeksu	280
10.2. Agregacja danych	281
Przetwarzanie kolumna po kolumnie i stosowanie wielu funkcji	282
Zwracanie zagregowanych danych bez indeksów wierszy	285
10.3. Metoda apply — ogólne zastosowanie techniki dziel-zastosuj-połącz	286
Usuwanie kluczy grup	288
Kwantyle i analiza koszykowa	288
Przykład: wypełnianie brakujących wartości przy użyciu wartości charakterystycznych dla grupy	290
Przykład: losowe generowanie próbek i permutacja	292
Przykład: średnie ważone grup i współczynnik korelacji	293
Przykład: regresja liniowa grup	295
10.4. Tabele przestawne i krzyżowe	295
Tabele krzyżowe	298
10.5. Podsumowanie	299
11. Szeregi czasowe	301
11.1. Typy danych i narzędzia przeznaczone do obsługi daty i czasu	302
Konwersja pomiędzy obiektami string i datetime	303
11.2. Podstawy szeregów czasowych	305
Indeksowanie i wybieranie	306
Szeregi czasowe z duplikatami indeksów	309
11.3. Zakresy dat, częstotliwości i przesunięcia	310
Generowanie zakresów dat	310
Częstotliwości i przesunięcia daty	313
Przesuwanie daty	314
11.4. Obsługa strefy czasowej	317
Lokalizacja i konwersja stref czasowych	317
Operacje z udziałem obiektów Timestamp o wyznaczonej strefie czasowej	319
Operacje pomiędzy różnymi strefami czasowymi	320
11.5. Okresy i przeprowadzanie na nich operacji matematycznych	321
Konwersja częstotliwości łańcuchów	322
Kwartalne częstotliwości okresów	323
Konwersja znaczników czasu na okresy (i z powrotem)	325
Tworzenie obiektów Periodindex na podstawie tablic	326

11.6. Zmiana rozdzielczości i konwersja częstotliwości	328
Zmniejszanie częstotliwości	329
Zwiększanie rozdzielczości i interpolacja	332
Zmiana rozdzielczości z okresami	333
11.7. Funkcje ruchomego okna	334
Funkcje ważone wykładniczo	337
Binarne funkcje ruchomego okna	338
Funkcje ruchomego okna definiowane przez użytkownika	340
11.8. Podsumowanie	340
12. Zaawansowane funkcje biblioteki pandas	341
12.1. Dane kateryczne	341
Kontekst i motywacja	341
Typ Categorical w bibliotece pandas	343
Obliczenia na obiektach typu Categorical	345
Metody obiektu katerycznego	347
12.2. Zaawansowane operacje grupowania	349
Transformacje grup i „nieobudowane” operacje grupowania	349
Zmiana rozdzielczości czasu przeprowadzana przy użyciu grup	353
12.3. Techniki łączenia metod w łańcuch	354
Metoda pipe	355
12.4. Podsumowanie	356
13. Wprowadzenie do bibliotek modelujących	357
13.1. Łączenie pandas z kodem modelu	357
13.2. Tworzenie opisów modeli przy użyciu biblioteki Patsy	360
Przekształcenia danych za pomocą formuł Patsy	362
Patsy i dane kateryczne	363
13.3. Wprowadzenie do biblioteki statsmodels	366
Szacowanie modeli liniowych	366
Szacowanie procesów szeregów czasowych	369
13.4. Wprowadzenie do pakietu scikit-learn	369
13.5. Dalszy rozwój	373
14. Przykłady analizy danych	375
14.1. Dane USA.gov serwisu Bitly	375
Liczenie stref czasowych w czystym Pythonie	376
Liczenie stref czasowych przy użyciu pakietu pandas	378
14.2. Zbiór danych MovieLens 1M	384
Wyznaczenie rozbieżności ocen	388
14.3. Imiona nadawane dzieciom w USA w latach 1880 - 2010	389
Analiza trendów imion	394
14.4. Baza danych USDA Food	402
14.5. Baza danych 2012 Federal Election Commission	406
Statystyki datków z podziałem na wykonywany zawód i pracodawcę	409
Podział kwot datków na koszyki	411

Statystyki datków z podziałem na poszczególne stany	413
14.6. Podsumowanie	414
A. Zaawansowane zagadnienia związane z biblioteką NumPy	415
A. 1. Szczegóły budowy obiektu ndarray	415
Hierarchia typów danych NumPy	416
A.2. Zaawansowane operacje tablicowe	417
Zmiana wymiarów tablic	417
Kolejności charakterystyczne dla języków C i Fortran	419
Łączenie i dzielenie tablic	420
Powtarzanie elementów — funkcje tile i repeat	422
Alternatywy indeksowania specjalnego — metody take i put	423
A.3. Rozgłaszanie	424
Rozgłaszanie wzdłuż innych osi	426
Przypisywanie wartości elementom tablicy poprzez rozgłaszanie	428
A.4. Zaawansowane zastosowania funkcji uniwersalnych	429
Metody instancji funkcji uniwersalnych	429
Pisanie nowych funkcji uniwersalnych w Pythonie	431
A.5. Tablice o złożonej strukturze	432
Zagnieżdżone typy danych i pola wielowymiarowe	433
Do czego przydają się tablice o złożonej strukturze?	434
A.6. Jeszcze coś o sortowaniu	434
Sortowanie pośrednie — metody argsort i lexsort	435
Alternatywne algorytmy sortowania	436
Częściowe sortowanie tablic	437
Wyszukiwanie elementów w posortowanej tablicy za pomocą metody numpy.searchsorted	438
A.7. Pisanie szybkich funkcji NumPy za pomocą pakietu Numba	439
Tworzenie obiektów numpy.ufunc za pomocą pakietu Numba	440
A.8. Zaawansowane tablicowe operacje wejścia i wyjścia	441
Pliki mapowane w pamięci	441
HDF5 i inne możliwości zapisu tablic	442
A.9. Jak zachować wysoką wydajność?	442
Dlaczego warto korzystać z sąsiadujących ze sobą obszarów pamięci?	443
B. Dodatkowe informacje dotyczące systemu IPython	445
B.1. Korzystanie z historii poleceń	445
Przeszukiwanie i korzystanie z historii poleceń	445
Zmienne wejściowe i wyjściowe	446
B.2. Interakcja z systemem operacyjnym	447
Polecenia powłoki systemowej i aliasy	447
System tworzenia skrótów do katalogów	448
B.3. Narzędzia programistyczne	449
Interaktywny debugger	449
Pomiar czasu — funkcje %time i %timeit	453
Podstawowe profilowanie — funkcje %prun i %run-p	455

Profilowanie funkcji linia po linii	457
B.4. Wskazówki dotyczące produktywnego tworzenia kodu w środowisku IPython	458
Przeładowywanie modułów	459
Wskazówki dotyczące projektowania kodu	460
B.5. Zaawansowane funkcje środowiska IPython	461
Co zrobić, aby własne klasy były przyjazne dla systemu IPython?	461
Profile i konfiguracja	462
Skorowidz	465

oprac. BPK