

# Przetwarzanie i analiza danych w języku Python

Celem autorów książki jest przygotowanie Czytelnika do samodzielnego przeprowadzenia całego procesu analizy danych, od pobrania i załadowania zbioru, przez jego wstępne przetworzenie i wyczyszczenie, aż po samą analizę, wizualizację wyników i ich interpretację. Wiemy, że pewne rozwiązania, które stworzy Czytelnik, przeznaczone będą do wielokrotnego użytku i tym samym zasługiwać będą na wdrożenie w ramach większych projektów informatycznych. Z tego powodu omawiamy także zestaw dobrych praktyk inżynierii oprogramowania.

Publikacja zawiera szereg przykładów, od prostych do bardziej rozbudowanych, pozwalających zrozumieć nie tylko poszczególne etapy procesu analizy danych, ale również zasady funkcjonowania środowiska Python 3. Czytelna struktura książki umożliwia osobom mającym już pewną wiedzę łatwe wyszukanie tylko wybranych, interesujących ich zagadnień.

*Przetwarzanie i analiza danych w języku Python* jest podsumowaniem doświadczeń autorów wyniesionych z zajęć prowadzonych na Wydziale Matematyki i Nauk Informatycznych Politechniki Warszawskiej (m.in. dla studentów matematyki i informatyki ze specjalności dotyczących statystyki matematycznej, analizy danych i *data science*), licznych szkoleń (np. dla *Data Science Retreat* w Berlinie), a także z pracy naukowo-badawczej w Instytucie Badań Systemowych Polskiej Akademii Nauk oraz w Instytucie Podstaw Informatyki PAN w ramach *International Ph.D. Studies Program* (m. in. w dziedzinie analizy i agregacji danych).



Wydawnictwo  
Naukowe PWN SA  
pwn.pl • 801 33 33 88  
ksiegarnia.pwn.pl



Marek Gągolewski  
Maciej Bartoszek  
Anna Cena

Przetwarzanie i analiza danych w języku Python



Marek Gągolewski  
Maciej Bartoszek  
Anna Cena

# Przetwarzanie i analiza danych w języku Python



# Przetwarzanie i analiza danych w języku Python

Marek Gągolewski  
Maciej Bartoszek  
Anna Cena

# Przetwarzanie i analiza danych w języku Python



# SPIS TREŚCI

---

Przedmowa . . . . .	XI
---------------------	----

---

## I Podstawy języka Python

<b>1. Wprowadzenie . . . . .</b>	<b>3</b>
1.1. Język i środowisko Python . . . . .	3
1.1.1. Instalacja dystrybucji środowiska Python . . . . .	3
1.1.2. Instalacja pakietów . . . . .	5
1.2. Notatniki Jupyter . . . . .	7
1.2.1. Tryby pracy . . . . .	7
1.2.2. Najważniejsze skróty klawiszowe . . . . .	10
1.2.3. Podstawy języka Markdown . . . . .	10
1.3. Pierwsze kroki w języku Python . . . . .	12
<b>2. Typy skalarne . . . . .</b>	<b>16</b>
2.1. Liczby . . . . .	16
2.1.1. Operatory arytmetyczne . . . . .	18
2.1.2. Konwersja typów . . . . .	21
2.1.3. Tworzenie obiektów nazwanych . . . . .	22
2.1.4. Funkcje wbudowane . . . . .	23
2.1.5. Pola i metody . . . . .	24
2.1.6. Arytmetyka zmiennopozycyjna . . . . .	25
2.2. Wartości logiczne . . . . .	26
2.2.1. Operatory relacyjne . . . . .	27
2.2.2. Operatory logiczne . . . . .	28
2.3. Napisy . . . . .	28
2.3.1. Tworzenie napisów . . . . .	28
2.3.2. Podstawowe operacje na napisach . . . . .	30
<b>3. Typy sekwencyjne i iterowalne . . . . .</b>	<b>32</b>
3.1. Podstawowe rodziny obiektów typu sekwencyjnego . . . . .	33
3.1.1. Listy i krotki . . . . .	33
3.1.2. Zakresy . . . . .	35
3.1.3. Napisy . . . . .	35

3.2.	Zarządzanie elementami . . . . .	35
3.2.1.	Wybieranie elementów . . . . .	35
3.2.2.	Modyfikacja elementów . . . . .	38
3.2.3.	Dodawanie i usuwanie elementów . . . . .	39
3.2.4.	Kopiowanie referencji, kopiowanie płytkie a głębokie . . . . .	41
3.3.	Obiekty iterowalne . . . . .	45
3.4.	Działania na obiektach iterowalnych i typu sekwencyjnego . . . . .	47
3.4.1.	Podstawowe metody i funkcje . . . . .	47
3.4.2.	Krotki identyfikatorów po lewej stronie operatora przypisania . . . . .	50
3.4.3.	Wyrażenia listotwórcze i generatory . . . . .	51
3.4.4.	Formatowanie napisów . . . . .	54
<b>4.</b>	<b>Słowniki i zbiory . . . . .</b>	<b>56</b>
4.1.	Słowniki . . . . .	56
4.1.1.	Tworzenie słowników . . . . .	56
4.1.2.	Podstawowe metody i funkcje . . . . .	58
4.2.	Zbiory . . . . .	61
4.2.1.	Tworzenie zbiorów . . . . .	61
4.2.2.	Podstawowe metody i funkcje . . . . .	62
<b>5.</b>	<b>Instrukcje sterujące . . . . .</b>	<b>64</b>
5.1.	Instrukcja warunkowa . . . . .	64
5.2.	Pętle . . . . .	66
5.2.1.	Pętla <code>while</code> . . . . .	66
5.2.2.	Pętla <code>for</code> . . . . .	67
5.2.3.	Instrukcje <code>break</code> i <code>continue</code> oraz blok <code>else</code> w pętlach . . . . .	69
5.3.	Obsługa wyjątków . . . . .	73
5.3.1.	Zgłaszanie wyjątków . . . . .	74
5.3.2.	Rodzaje wyjątków . . . . .	74
5.3.3.	Wychwytywanie wyjątków . . . . .	75
<b>6.</b>	<b>Funkcje . . . . .</b>	<b>77</b>
6.1.	Definiowanie funkcji . . . . .	77
6.1.1.	Dokumentowanie funkcji . . . . .	78
6.1.2.	Wartość zwracana . . . . .	79
6.1.3.	Wyrażenia <code>lambda</code> . . . . .	80
6.2.	Parametry i argumenty . . . . .	81
6.2.1.	Sposób przekazywania argumentów . . . . .	81
6.2.2.	Sprawdzanie poprawności argumentów . . . . .	82
6.2.3.	Dopasowywanie argumentów . . . . .	84
6.2.4.	Parametry z argumentami domyślnymi . . . . .	84
6.2.5.	Rozpakowywanie argumentów . . . . .	85
6.2.6.	Parametry specjalne <code>*args</code> i <code>**kwargs</code> . . . . .	86
6.3.	Zasięg zmiennych . . . . .	88
6.3.1.	Zmienne lokalne . . . . .	88
6.3.2.	Zmienne globalne . . . . .	88
6.3.3.	Zmienne nielokalne, fabryki funkcji i domknięcia . . . . .	90
6.4.	Pakiety . . . . .	92

## II Przetwarzanie danych

<b>7. Wektory, macierze i inne tablice</b>	97
7.1. Tworzenie i reprezentacja tablic	97
7.1.1. Funkcja <code>array()</code>	98
7.1.2. Reprezentacja tablic	100
7.1.3. Typ przechowywanych elementów	101
7.1.4. Tworzenie tablic specjalnego rodzaju	103
7.1.5. Łączenie tablic	106
7.2. Podstawowe metody i funkcje	108
7.2.1. Operatory arytmetyczne. Uzgadnianie kształtów	108
7.2.2. Operacje relacyjne i logiczne	113
7.2.3. Zwektoryzowane funkcje matematyczne	115
7.2.4. Agregacja danych	118
7.2.5. Inne operacje	121
7.3. Indeksowanie tablic	123
7.3.1. Indeksowanie wektorów	123
7.3.2. Indeksowanie macierzy	128
7.3.3. Indeksowanie tablic $N$ -wymiarowych	132
7.3.4. Wyszukiwanie indeksów elementów spełniających zadane kryteria	134
<b>8. Ramki danych</b>	137
8.1. Tworzenie ramek danych	138
8.1.1. Konstruktor klasy <code>DataFrame</code>	138
8.1.2. Importowanie ramek danych z plików i innych źródeł	139
8.1.3. Odczytywanie podstawowych informacji o ramkach danych	140
8.2. Zmienne, czyli obiekty typu <code>Series</code>	143
8.2.1. Wydobywanie poszczególnych zmiennych	143
8.2.2. Tworzenie i reprezentacja zmiennych	144
8.2.3. Zmienne typu <code>data</code> i <code>czas</code>	145
8.2.4. Zmienne jakościowe i porządkowe	146
8.3. Etykiety, czyli obiekty typu <code>Index</code>	150
8.3.1. Etykietowanie wierszy i kolumn	151
8.3.2. Etykiety hierarchiczne	152
8.4. Indeksowanie zmiennych i ramek danych	154
8.4.1. Wybór elementów pojedynczej zmiennej	154
8.4.2. Wybór podzbioru wierszy i kolumn ramki danych	160
8.5. Wybrane operacje	164
8.5.1. Dodawanie oraz usuwanie kolumn i wierszy	164
8.5.2. Przekształcanie zmiennych	166
8.5.3. Podsumowania ramek danych i zmiennych	168
8.5.4. Sortowanie ramek danych	172
8.5.5. Zmiana kształtu ramek danych	173
8.5.6. Obserwacje brakujące	176
<b>9. Przetwarzanie napisów</b>	179
9.1. Operacje na pojedynczych napisach	179
9.1.1. Podstawowe stałe napisowe i operacje na pojedynczych znakach	180

9.1.2.	Wyszukiwanie ustalonego wzorca . . . . .	182
9.1.3.	Translacja znaków . . . . .	183
9.1.4.	Sprawdzanie, czy wszystkie znaki należą do podanej kategorii . . . . .	184
9.1.5.	Dzielenie i sklejanie tekstu . . . . .	184
9.2.	Wyszukiwanie wzorca przy użyciu wyrażeń regularnych . . . . .	185
9.2.1.	Definiowanie wyrażeń regularnych . . . . .	186
9.2.2.	Przegląd funkcji . . . . .	188
9.2.3.	Wydzielone podwyrażenia i odwołania do nich . . . . .	189
9.3.	Zwektoryzowane operacje na obiektach <code>Index</code> i <code>Series</code> . . . . .	190
<b>10.</b>	<b>Przetwarzanie plików i zasobów w internecie . . . . .</b>	<b>196</b>
10.1.	Operacje na drzewie katalogów . . . . .	196
10.1.1.	Ścieżki dostępu . . . . .	196
10.1.2.	Wyszukiwanie plików na dysku . . . . .	198
10.2.	Przetwarzanie plików . . . . .	200
10.2.1.	Otwieranie pliku w różnych trybach . . . . .	200
10.2.2.	Odczytywanie zawartości pliku . . . . .	202
10.2.3.	Zapisywanie danych do pliku . . . . .	203
10.2.4.	Serializacja obiektów . . . . .	204
10.2.5.	Popularne formaty plików . . . . .	205
10.3.	Pozyskiwanie danych ze stron internetowych . . . . .	208
10.3.1.	Wydobywanie tabel w postaci ramek danych . . . . .	209
10.3.2.	Ręczne przetwarzanie kodu źródłowego strony . . . . .	209
10.3.3.	Parsowanie kodu HTML i wydobywanie pojedynczych elementów . . . . .	211
<b>11.</b>	<b>Dostęp do baz danych . . . . .</b>	<b>215</b>
11.1.	Przykładowa baza danych: <code>nycflights13</code> . . . . .	215
11.2.	Obsługa baz danych . . . . .	218
11.2.1.	Połączenie z bazą danych . . . . .	218
11.2.2.	Eksportowanie danych do bazy . . . . .	218
11.2.3.	Odczytywanie danych z bazy . . . . .	219
11.2.4.	Funkcje z pakietu <code>pandas</code> . . . . .	220
11.3.	Ćwiczenia . . . . .	221
11.3.1.	Wybór unikatowych podzbiorów kolumn . . . . .	222
11.3.2.	Agregacja danych w podgrupach . . . . .	223
11.3.3.	Filtrowanie danych wejściowych i wyników . . . . .	226
11.3.4.	Sortowanie wyników . . . . .	230
11.3.5.	Operacje teorii mnogościowe . . . . .	232
11.3.6.	Złączenia . . . . .	234

---

### III Analiza danych

<b>12.</b>	<b>Wizualizacja danych . . . . .</b>	<b>239</b>
12.1.	Rysowanie podstawowych obiektów . . . . .	240
12.1.1.	Łamane . . . . .	240
12.1.2.	Punkty i różne symbole . . . . .	241
12.1.3.	Wielokąty . . . . .	242
12.1.4.	Adnotacje tekstowe . . . . .	243

12.2. Parametry graficzne . . . . .	244
12.2.1. Sposoby kreślenia punktów i odcinków . . . . .	244
12.2.2. Sposoby określania barw . . . . .	244
12.2.3. Napisy formatujące . . . . .	246
12.2.4. Ustawienia osi . . . . .	247
12.3. Rysunki jako kombinacje obiektów podstawowych . . . . .	248
12.3.1. Wiele obiektów na jednym wykresie . . . . .	248
12.3.2. Legenda . . . . .	250
12.3.3. Wiele wykresów na jednej stronie . . . . .	251
12.4. Graficzna prezentacja danych . . . . .	255
12.4.1. Wybrane wykresy dla danych jakościowych . . . . .	255
12.4.2. Wybrane wykresy dla danych ilościowych . . . . .	258
12.4.3. Wybrane wykresy dla funkcji dwuwymiarowych . . . . .	262
<b>13. Wnioskowanie statystyczne . . . . .</b>	<b>265</b>
13.1. Wybrane rozkłady prawdopodobieństwa . . . . .	265
13.1.1. Podstawowe rodziny rozkładów . . . . .	265
13.1.2. Generowanie liczb pseudolosowych . . . . .	273
13.2. Estymacja parametrów i charakterystyk rozkładów . . . . .	275
13.2.1. Estymacja punktowa . . . . .	276
13.2.2. Estymacja przedziałowa . . . . .	278
13.3. Wykorzystanie testów statystycznych w analizie danych . . . . .	280
13.3.1. Testy zgodności . . . . .	281
13.3.2. Testy parametryczne . . . . .	290
13.3.3. Testy nieparametryczne . . . . .	295
<b>14. Wybrane algorytmy uczenia maszynowego . . . . .</b>	<b>298</b>
14.1. Przykładowy zbiór danych: winequality . . . . .	298
14.2. Analiza regresji . . . . .	300
14.2.1. Regresja liniowa . . . . .	301
14.2.2. Ocena jakości dopasowania modelu . . . . .	304
14.2.3. Model wielomianowy . . . . .	306
14.2.4. Wybór zmiennych do modelu . . . . .	307
14.3. Klasyfikacja . . . . .	310
14.3.1. Metoda $k$ -najbliższych sąsiadów . . . . .	312
14.3.2. Ocena jakości klasyfikatora . . . . .	312
14.3.3. Drzewa decyzyjne i lasy losowe . . . . .	315
14.3.4. Porównanie krzyżowe . . . . .	318
14.4. Analiza skupień . . . . .	320
14.4.1. Algorytm $k$ -średnich . . . . .	320
14.4.2. Hierarchiczna analiza skupień . . . . .	326
<hr/>	
<b>IV Tworzenie własnego oprogramowania</b>	
<b>15. Moduły, pakiety i skrypty . . . . .</b>	<b>331</b>
15.1. Projekty wielomodułowe . . . . .	331
15.1.1. Środowisko programistyczne Spyder . . . . .	331
15.1.2. Tworzenie i ładowanie modułów . . . . .	332



15.1.3. Tworzenie i ładowanie pakietów . . . . .	335
15.1.4. Ścieżki wyszukiwania modułów i pakietów . . . . .	336
15.2. Skrypty . . . . .	336
15.2.1. Uruchomienie skryptu z poziomu powłoki . . . . .	337
15.2.2. Przekazywanie argumentów . . . . .	338
15.2.3. Skrypty a moduły. Testy jednostkowe . . . . .	339
<b>16. Programowanie obiektowe . . . . .</b>	<b>343</b>
16.1. Klasy i relacje między nimi . . . . .	344
16.1.1. Definiowanie klasy . . . . .	344
16.1.2. Dziedziczenie . . . . .	346
16.2. Metody . . . . .	348
16.2.1. Przeciążanie metod. Polimorfizm . . . . .	348
16.2.2. Metody i pola statyczne . . . . .	350
16.2.3. Metody specjalne . . . . .	351
16.3. Pola . . . . .	357
16.3.1. Definiowanie z góry ustalonych pól w klasie . . . . .	357
16.3.2. Pola prywatne, chronione i publiczne . . . . .	358
<b>Bibliografia . . . . .</b>	<b>361</b>
<b>Skorowidz . . . . .</b>	<b>363</b>