

---

# Spis treści

Przedmowa .....	XV
-----------------	----

---

## Część I. Danetyczny cykl życia

<b>1. Danetyczny cykl życia .....</b>	<b>3</b>
Etapy cyklu życia .....	3
Przykłady cyklu życia .....	6
Podsumowanie .....	7
<b>2. Pytania i zakres danych .....</b>	<b>9</b>
Big Data i nowe możliwości .....	10
Przykład: system Google do śledzenia zachorowań na grypę .....	10
Populacja docelowa, zbiór dostępny i próba .....	12
Przykład: co sprawia, że członkowie społeczności online są aktywni? .....	14
Przykład: kto wygra wybory? .....	15
Przykład: jaki jest związek zagrożeń środowiskowych ze zdrowiem osób? .....	16
Przyrządy i protokoły .....	17
Mierzenie zjawiska naturalnego .....	17
Przykład: jaki jest poziom CO <sub>2</sub> w powietrzu? .....	18
Dokładność .....	19
Typy obciążenia .....	21
Typy wariacji .....	23
Podsumowanie .....	25
<b>3. Symulacja i projekt danych .....</b>	<b>27</b>
Model urnowy .....	28
Projekty próbkowania .....	30

Próbkowanie rozkładu statystycznego .....	32
Symulacja rozkładu próbkowania .....	34
Symulacja z rozkładem hipergeometrycznym.....	35
Przykład: symulowanie obciążenia systematycznego i wariacji sondażu	
wyborczego.....	37
Model urnowy dla Pensylwanii .....	38
Model urnowy z obciążeniem .....	41
Prowadzenie większych sondaży .....	42
Przykład: symulacja randomizowanego badania klinicznego szczepionki .....	44
Zakres .....	44
Model urnowy dla przypisania losowego .....	45
Przykład: pomiary jakości powietrza .....	47
Podsumowanie .....	50
<b>4. Modelowanie przy użyciu statystyk podsumowujących .....</b>	<b>53</b>
Model stałej.....	54
Minimalizacja straty .....	56
Średni błąd bezwzględny .....	57
Średni błąd kwadratowy .....	59
Wybór funkcji straty .....	62
Podsumowanie .....	62
<b>5. Studium przypadku: dlaczego mój autobus zawsze się spóźnia? .....</b>	<b>65</b>
Pytanie i zakres.....	66
Przetwarzanie danych .....	66
Eksplorowanie czasów autobusów.....	69
Modelowanie czasów oczekiwania .....	72
Podsumowanie .....	77

---

## Część II. Dane prostokątne

<b>6. Praca z ramkami danych przy użyciu pandas.....</b>	<b>81</b>
Operacje na podzbiorach .....	82
Zakres danych i pytanie .....	82
Ramki danych i indeksy .....	84
Wycinanie.....	85
Filtrowanie wierszy .....	89
Przykład: kiedy imię Luna stało się popularne? .....	92
Agregacje.....	94

Podstawowe grupowanie-agregowanie . . . . .	95
Grupowanie po wielu kolumnach . . . . .	98
Niestandardowe funkcje agregujące . . . . .	99
Przestawianie . . . . .	102
Złączenia . . . . .	104
Złączenia wewnętrzne . . . . .	106
Złączenia lewostronne, prawostronne i zewnętrzne . . . . .	107
Przykład: popularność kategorii imion NYT . . . . .	109
Przekształcanie . . . . .	111
Metoda apply . . . . .	112
Przykład: popularność imion na literę „L” . . . . .	114
Cena metody apply . . . . .	115
Czym różnią się ramki danych od innych reprezentacji danych? . . . . .	116
Ramki danych a arkusze kalkulacyjne . . . . .	116
Ramki danych a macierze . . . . .	117
Ramki danych a relacje . . . . .	118
Podsumowanie . . . . .	118
<b>7. Praca z relacjami przy użyciu SQL . . . . .</b>	<b>119</b>
Operacje na podzbiorach . . . . .	119
Podstawy SQL: SELECT i FROM . . . . .	120
Czym jest relacja? . . . . .	121
Wycinanie . . . . .	122
Filtrowanie wierszy . . . . .	124
Przykład: kiedy imię Luna stało się popularne? . . . . .	126
Agregacje . . . . .	127
Podstawowe grupowanie-agregowanie przy użyciu GROUP BY . . . . .	128
Grupowanie po wielu kolumnach . . . . .	130
Inne funkcje agregujące . . . . .	130
Złączenia . . . . .	131
Złączenia wewnętrzne . . . . .	133
Złączenia lewostronne i prawostronne . . . . .	134
Przykład: popularność kategorii imion NYT . . . . .	136
Przekształcanie i wyrażenia CTE . . . . .	137
Funkcje SQL . . . . .	137
Zapytania wieloetapowe przy użyciu klauzuli WITH . . . . .	140
Przykład: popularność imion na literę „L” . . . . .	141
Podsumowanie . . . . .	142

---

## Część III. Interpretacja danych

<b>8. Przekształcanie plików</b> .....	<b>145</b>
Przykładowe źródła danych .....	146
Ankieta DAWN .....	146
Bezpieczeństwo żywności w restauracjach w San Francisco .....	147
Formaty plików .....	148
Format rozdzielany .....	148
Format o stałej szerokości .....	150
Formaty hierarchiczne .....	151
Swobodnie sformatowany tekst .....	152
Kodowanie plików .....	152
Rozmiar pliku .....	154
Powłoka i narzędzia wiersza poleceń .....	158
Kształt i ziarnistość tabeli .....	162
Ziarnistość kontroli i naruszeń w restauracjach .....	164
Kształt i ziarnistość ankiety DAWN .....	166
Podsumowanie .....	169
<b>9. Przekształcanie ramek danych</b> .....	<b>171</b>
Przykład: przekształcanie pomiarów CO <sub>2</sub> z obserwatorium Mauna Loa ....	172
Testy jakości .....	175
Obsługa brakujących danych .....	178
Zmiana kształtu tabeli danych .....	179
Testy jakości .....	180
Jakość na podstawie zakresu .....	180
Jakość pomiarów i zarejestrowanych wartości .....	181
Jakość wśród powiązanych cech .....	182
Jakość analizy .....	183
Poprawiać dane czy nie .....	183
Brakujące wartości i rekordy .....	185
Transformacje i znaczniki czasu .....	187
Transformacje znaczników czasu .....	188
Tworzenie potoków transformacji .....	191
Modyfikowanie struktury .....	192
Przykład: przekształcanie naruszeń bezpieczeństwa restauracji .....	195
Zawężenie skupienia .....	196
Agregowanie naruszeń .....	197
Wyodrębnianie informacji z opisów naruszeń .....	199

Podsumowanie .....	203
<b>10. Eksploracyjna analiza danych .....</b>	<b>205</b>
Typy cech .....	207
Przykład: rasy psów .....	208
Transformowanie cech jakościowych .....	215
Znaczenie typów cech .....	218
Czego szukać w rozkładzie .....	219
Czego szukać w zależności .....	223
Dwie cechy ilościowe .....	224
Jedna zmienna jakościowa i jedna zmienna ilościowa .....	225
Dwie cechy jakościowe .....	227
Porównania w sytuacjach wielu zmiennych .....	228
Wytyczne dotyczące eksploracji .....	232
Przykład: ceny sprzedaży domów .....	233
Interpretacja ceny .....	234
Co dalej? .....	237
Badanie innych cech .....	238
Głębsze poszukiwanie zależności .....	242
Ustalanie lokalizacji .....	244
Odkrycia eksploracyjnej analizy danych .....	246
Podsumowanie .....	246
<b>11. Wizualizacja danych .....</b>	<b>249</b>
Wybór skali, aby ujawnić strukturę .....	249
Wypełnianie obszaru danych .....	250
Uwzględnianie zera .....	251
Ujawnianie kształtu przez transformacje .....	253
Przechyłanie w celu rozszyfrowania zależności .....	256
Odkrywanie zależności przez prostowanie .....	256
Wygładzanie i agregowanie danych .....	259
Techniki wygładzania, aby odkryć kształt .....	260
Techniki wygładzania, aby odkryć zależności i trendy .....	261
Techniki wygładzania wymagają dostrojenia .....	264
Redukcja rozkładów do kwantyli .....	265
Kiedy nie wygładzać .....	267
Ułatwianie znaczących porównań .....	269
Podkreślanie istotniej różnicy .....	269
Porządkowanie grup .....	271
Unikanie kumulowania .....	273

Wybór palety kolorów.....	275
Wytyczne dotyczące porównań na wykresach.....	277
Wykorzystanie projektu danych.....	278
Dane zbierane w czasie.....	278
Badania obserwacyjne.....	280
Nierówne próbkowanie.....	281
Dane geograficzne.....	282
Dodawanie kontekstu.....	283
Przykład: czasy sprintu na 100 m.....	283
Tworzenie wykresów przy użyciu plotly.....	285
Obiekty Figure i Trace.....	286
Modyfikowanie układu.....	287
Funkcje kreślenia.....	289
Adnotacje.....	292
Inne narzędzia do wizualizacji.....	293
matplotlib.....	293
Gramatyka grafiki.....	293
Podsumowanie.....	294
<b>12. Studium przypadku: jak dokładne są pomiary jakości powietrza?.....</b>	<b>295</b>
Pytanie, projekt i zakres.....	297
Znajdowanie kolokowanych czujników.....	298
Przekształcanie listy lokalizacji AQS.....	299
Przekształcanie listy lokalizacji PurpleAir.....	301
Dopasowywanie czujników AQS i PurpleAir.....	303
Przekształcanie i oczyszczanie danych czujnika AQS.....	305
Sprawdzanie ziarnistości.....	306
Usuwanie zbędnych kolumn.....	308
Sprawdzanie poprawności dat.....	308
Sprawdzanie jakości pomiarów PM <sub>2,5</sub> .....	309
Przekształcanie danych czujnika PurpleAir.....	310
Sprawdzanie ziarnistości.....	312
Obsługa brakujących wartości.....	317
Eksplorowanie pomiarów PurpleAir i AQS.....	319
Tworzenie modelu do korygowania pomiarów PurpleAir.....	325
Podsumowanie.....	328

---

## Część IV. Inne źródła danych

<b>13. Praca z tekstem</b> .....	<b>333</b>
Przykłady tekstu i zadań .....	334
Konwersja tekstu na format standardowy .....	334
Wyodrębnianie fragmentu tekstu, aby utworzyć cechę .....	334
Transformacja tekstu na cechy .....	335
Analiza tekstu .....	335
Manipulacja ciągami znaków .....	336
Konwertowanie tekstu na format standardowy przy użyciu metod ciągów znaków Pythona .....	336
Metody ciągów znaków w pandas .....	338
Rozdzielanie ciągów w celu wyodrębnienia fragmentów tekstu .....	339
Wyrażenia regularne .....	340
Konkatenacja literałów .....	340
Kwantyfikatory .....	344
Alternatywa i grupowanie w celu tworzenia cech .....	345
Tabele referencyjne .....	346
Analiza tekstu .....	350
Podsumowanie .....	355
<b>14. Wymiana danych</b> .....	<b>357</b>
Dane NetCDF .....	358
Dane JSON .....	364
HTTP .....	369
REST .....	373
XML, HTML i XPath .....	378
Przykład: pozyskiwanie czasów wyścigu z Wikipedii .....	381
XPath .....	384
Przykład: dostęp do kursów wymiany z EBC .....	386
Podsumowanie .....	389

---

## Część V. Modelowanie liniowe

<b>15. Modele liniowe</b> .....	<b>393</b>
Prosty model liniowy .....	394
Przykład: prosty model liniowy dotyczący jakości powietrza .....	398
Interpretowanie modeli liniowych .....	401
Ocena dopasowania .....	402

Dopasowywanie prostego modelu liniowego . . . . .	403
Wielowymiarowy model liniowy . . . . .	405
Dopasowywanie wielowymiarowego modelu liniowego . . . . .	410
Przykład: gdzie leży kraina możliwości? . . . . .	414
Wyjaśnianie awansu ekonomicznego przy użyciu czasu dojazdów . . . . .	415
Związek wielu zmiennych z awansem ekonomicznym . . . . .	418
Inżynieria cech w przypadku pomiarów liczbowych . . . . .	423
Inżynieria cech w przypadku pomiarów kategoryalnych . . . . .	427
Podsumowanie . . . . .	435
<b>16. Wybór modelu . . . . .</b>	<b>437</b>
Nadmierne dopasowanie. . . . .	438
Przykład: zużycie energii . . . . .	438
Podział na zbiór uczący i testowy. . . . .	443
Walidacja krzyżowa . . . . .	448
Regularyzacja . . . . .	453
Obciążenie systematyczne i wariancja modelu . . . . .	454
Podsumowanie . . . . .	458
<b>17. Teoria wnioskowania i prognozowania . . . . .</b>	<b>461</b>
Rozkład: populacja, dane empiryczne, próbkowanie . . . . .	461
Podstawy testowania hipotez . . . . .	463
Przykład: test rankingowy porównujący produktywność współtwórców	
Wikipedii . . . . .	465
Przykład: test proporcji skuteczności szczepionki . . . . .	470
Stosowanie metody bootstrap do wnioskowania. . . . .	472
Podstawy przedziałów ufności . . . . .	477
Podstawy przedziałów prognoz . . . . .	480
Przykład: prognozowanie opóźnień autobusu. . . . .	481
Przykład: prognozowanie rozmiaru kraba . . . . .	482
Przykład: prognozowanie przyrostu kraba. . . . .	483
Prawdopodobieństwo wnioskowania i prognozowania . . . . .	486
Formalnie przedstawiona teoria statystyki średniej rankingowej. . . . .	486
Ogólne właściwości zmiennych losowych . . . . .	490
Prawdopodobieństwo dotyczące testowania i przedziałów . . . . .	492
Prawdopodobieństwo dotyczące wyboru modelu . . . . .	495
Podsumowanie . . . . .	497

<b>18. Studium przypadku: jak zważyć osła</b> .....	<b>499</b>
Pytanie i zakres badania dotyczącego osłów .....	499
Przekształcanie i transformacje .....	501
Eksploracja .....	506
Modelowanie ciężaru osła .....	509
Funkcja straty do przepisywania środków znieczulających .....	509
Dopasowywanie prostego modelu liniowego .....	510
Dopasowywanie wielowymiarowego modelu liniowego .....	512
Wprowadzenie cech jakościowych do modelu .....	513
Ocena modelu .....	517
Podsumowanie .....	519

---

## Część VI. Klasyfikacja

<b>19. Klasyfikacja</b> .....	<b>523</b>
Przykład: drzewa zniszczone przez wiatr .....	524
Modelowanie i klasyfikacja .....	526
Model stałej .....	526
Badanie zależności między rozmiarami a wykrotami .....	527
Modelowanie proporcji (i prawdopodobieństw) .....	530
Model logistyczny .....	531
Logarytm szansy .....	532
Stosowanie krzywej logistycznej .....	533
Funkcja straty dla modelu logistycznego .....	534
Od prawdopodobieństw do klasyfikacji .....	538
Macierz pomyłek .....	540
Precyzja kontra czułość .....	541
Podsumowanie .....	544
<b>20. Optymalizacja numeryczna</b> .....	<b>545</b>
Podstawy metody spadku gradientu .....	546
Minimalizacja straty Hubera .....	548
Wypukłe i różniczkowalne funkcje straty .....	551
Warianty spadku gradientu .....	552
Stochastyczny spadek gradientu .....	553
Miniwsadowy spadek gradientu .....	554
Metoda Newtona .....	554
Podsumowanie .....	555

<b>21. Studium przypadku: wykrywanie fałszywych wiadomości.....</b>	<b>557</b>
Pytanie i zakres.....	558
Pozyskiwanie i przekształcanie danych.....	559
Eksploracja danych .....	564
Eksplorowanie wydawców .....	565
Eksplorowanie daty publikacji.....	567
Eksplorowanie słów w artykułach .....	569
Modelowanie.....	571
Model z pojedynczym słowem.....	571
Model z wieloma słowami .....	573
Prognozowanie przy użyciu transformacji TFIDF .....	575
Podsumowanie .....	578
Bibliografia .....	581
Źródła danych .....	587
Indeks.....	593
O autorach .....	609