

Spis treści

O autorze	1
O recenzentach	1
Wprowadzenie	1
Rozdział 1. Oczyszczanie danych podczas importowania danych tabelarycznych do pandas	1
Wymagania techniczne	1
Importowanie danych z plików CSV	1
Przygotuj się	1
Jak to zrobić...	1
Jak to działa...	2
Zobacz również...	2
Co dalej?	2
Importowanie plików z Excela	2
Przygotuj się	2
Jak to zrobić...	2
Jak to działa...	2
Zobacz również...	2
Co dalej?	2
Importowanie danych z baz SQL	3
Przygotuj się	3
Jak to zrobić...	3
Jak to działa...	3
Zobacz również...	3
Co dalej?	3
Importowanie danych z SPSS, Stata i SAS	3
Przygotuj się	3
Jak to zrobić...	3

Jak to działa...	42
Zobacz również...	43
Co dalej?	43
Importowanie danych z R	43
Przygotuj się	44
Jak to zrobić...	44
Jak to działa...	46
Zobacz również...	47
Co dalej?	47
Przechowywanie danych tablicowych	48
Przygotuj się	49
Jak to zrobić...	49
Jak to działa...	51
Zobacz również	51
Rozdział 2. Oczyszczanie danych podczas importowania HTML-a i JSON-a do pandas	53
Wymagania techniczne	54
Importowanie danych z prostego pliku JSON	54
Przygotuj się	54
Jak to zrobić...	55
Jak to działa...	58
Zobacz również...	59
Importowanie bardziej złożonego JSON-a za pomocą API	60
Przygotuj się	60
Jak to zrobić...	61
Jak to działa...	63
Zobacz również...	64
Co dalej?	64
Importowanie danych ze stron internetowych	65
Przygotuj się	65
Jak to zrobić...	66
Jak to działa...	68
Zobacz również...	69
Przechowywanie danych w formacie JSON	69
Przygotuj się	70
Jak to zrobić...	71
Jak to działa...	72
Zobacz również...	73
Rozdział 3. Przeprowadzanie pomiarów danych	75
Wymagania techniczne	76
Pierwsze spojrzenie na dane	76
Przygotuj się...	77
Jak to zrobić...	77
Jak to działa...	79
Zobacz również...	80
Co dalej?	81

Wybór i organizacja kolumn	81
Przygotuj się...	81
Jak to zrobić...	81
Jak to działa...	85
Zobacz również...	86
Co dalej?	87
Selekcja wierszy	87
Przygotuj się...	87
Jak to zrobić...	87
Jak to działa...	94
Zobacz również...	95
Co dalej?	95
Obliczanie częstości zmiennych kategoryalnych	95
Przygotuj się...	95
Jak to zrobić...	95
Jak to działa...	98
Zobacz również...	99
Generowanie statystyk podsumowujących zmienne ciągłe	99
Przygotuj się...	100
Jak to zrobić...	100
Jak to działa...	102
Co dalej?	103
Rozdział 4. Identyfikacja brakujących i odstających wartości w podzbiorach danych	105
Wymagania techniczne	106
Wykrywanie brakujących wartości	106
Przygotuj się	106
Jak to zrobić...	107
Jak to działa...	109
Co dalej?	111
Identyfikowanie wartości odstających w pojedynczych zmiennych	111
Przygotuj się	111
Jak to zrobić...	111
Jak to działa...	117
Zobacz również...	117
Co dalej?	118
Identyfikacja wartości odstających i nieoczekiwanych w relacjach pomiędzy dwiema zmiennymi	111
Przygotuj się	111
Jak to zrobić...	119
Jak to działa...	122
Zobacz również...	125
Co dalej?	125
Wykorzystanie podzbiorów do badania logicznych niespójności w relacjach pomiędzy zmiennymi	125
Przygotuj się	125
Jak to zrobić...	127
Jak to działa...	128
Co dalej?	128

Wykorzystanie regresji liniowej do identyfikacji punktów danych o znaczącym wpływie	132
Przygotuj się	133
Jak to zrobić...	133
Jak to działa...	135
Zobacz również...	136
Znajdowanie wartości odstających za pomocą algorytmu k-najbliższych sąsiadów	136
Przygotuj się	136
Jak to zrobić...	137
Jak to działa...	138
Zobacz również...	139
Co dalej?	139
Wykorzystanie Isolation Forest do znajdowania anomalii	139
Przygotuj się	140
Jak to zrobić...	140
Jak to działa...	143
Zobacz również...	143
Co dalej?	143
Rozdział 5. Wykorzystanie wizualizacji do identyfikacji nieoczekiwanych wartości	145
Wymagania techniczne	146
Badanie rozkładu zmiennych ciągłych za pomocą histogramów	146
Przygotuj się	147
Jak to zrobić...	147
Jak to działa...	152
Zobacz również...	153
Identyfikacja wartości odstających w zmiennych ciągłych za pomocą wykresów pudełkowych	154
Przygotuj się	154
Jak to zrobić...	154
Jak to działa...	158
Zobacz również...	159
Co dalej?	159
Wykorzystanie grup wykresów pudełkowych do identyfikacji wartości nieoczekiwanych w określonej grupie	160
Przygotuj się	160
Jak to zrobić...	160
Jak to działa...	164
Zobacz również...	165
Co dalej?	166
Analiza wartości odstających i kształtu rozkładu za pomocą wykresów skrzypcowych	166
Przygotuj się	166
Jak to zrobić...	166
Jak to działa...	170
Zobacz również...	171
Co dalej?	172
Wykorzystanie wykresów punktowych do przedstawienia relacji dwuwymiarowych	172
Przygotuj się	172
Jak to zrobić...	173
Jak to działa...	178

Zobacz również...	17
Co dalej?	17
Wykorzystanie wykresów liniowych do analizy trendów zmiennych ciągłych	17
Przygotuj się	17
Jak to zrobić...	18
Jak to działa...	18
Zobacz również...	18
Co dalej?	18
Generowanie mapy ciepła na podstawie macierzy korelacji	18
Przygotuj się	18
Jak to zrobić...	18
Jak to działa...	18
Zobacz również...	18
Co dalej?	18
Rozdział 6. Oczyszczanie i eksploracja danych za pomocą operacji na obiektach typu Series	18
Wymagania techniczne	19
Pobieranie wartości z obiektów typu Series w pandas	19
Przygotuj się	19
Jak to zrobić...	19
Jak to działa...	19
Statystyki podsumowujące obiektów typu Series	14
Przygotuj się	15
Jak to zrobić...	15
Jak to działa...	17
Zobacz również...	18
Co dalej?	18
Zmiana wartości w obiektach typu Series	18
Przygotuj się	18
Jak to zrobić...	19
Jak to działa...	21
Zobacz również...	21
Co dalej?	22
Warunkowa zmiana wartości w obiektach typu Series	22
Przygotuj się	22
Jak to zrobić...	23
Jak to działa...	26
Zobacz również...	27
Co dalej?	28
Ocena zawartości i oczyszczanie serii łańcuchów znaków	28
Przygotuj się	28
Jak to zrobić...	28
Jak to działa...	22
Zobacz również...	22
Praca z datami	12
Przygotuj się	12
Jak to zrobić...	13
Jak to działa...	16
Co dalej?	17

Identyfikowanie i usuwanie braków w danych	217
Przygotuj się	218
Jak to zrobić...	218
Jak to działa...	221
Zobacz również...	221
Co dalej?	221
Imputacja brakujących wartości za pomocą metody k-najbliższych sąsiadów	222
Przygotuj się	222
Jak to zrobić...	222
Jak to działa...	223
Zobacz również...	223
Co dalej?	224
Rozdział 7. Porządkowanie danych podczas agregacji	225
Wymagania techniczne	226
Iteracje z użyciem itertuples (anty wzorzec)	226
Przygotuj się	227
Jak to zrobić...	227
Jak to działa...	229
Zobacz również...	230
Obliczanie statystyk podsumowujących poszczególne grupy za pomocą tablic NumPy	231
Przygotuj się	231
Jak to zrobić...	231
Jak to działa...	233
Zobacz również...	233
Co dalej?	233
Grupowanie danych za pomocą groupby	234
Przygotuj się	234
Jak to zrobić...	234
Jak to działa...	236
Zobacz również...	236
Korzystanie z bardziej skomplikowanych funkcji agregujących i groupby	237
Przygotuj się	237
Jak to zrobić...	237
Jak to działa...	240
Zobacz również...	241
Co dalej?	242
groupby i funkcje zdefiniowane przez użytkownika	242
Przygotuj się	242
Jak to zrobić...	242
Jak to działa...	245
Zobacz również...	245
Co dalej?	246
Wykorzystanie groupby do zmiany jednostki analizy w ramce	246
Przygotuj się	246
Jak to zrobić...	246
Jak to działa...	247

Rozdział 8. Rozwiązywanie problemów z danymi podczas łączenia ramek danych	249
Wymagania techniczne	250
Łączenie ramek danych w pionie	250
Przygotuj się	251
Jak to zrobić...	251
Jak to działa...	253
Co dalej?	254
Wykonywanie połączeń jeden-do-jednego	254
Przygotuj się	256
Jak to zrobić...	256
Jak to działa...	259
Zobacz również...	260
Scalania w wielu kolumnach	260
Przygotuj się	260
Jak to zrobić...	261
Jak to działa...	262
Zobacz również...	263
Wykonywanie połączeń jeden-do-wielu	263
Przygotuj się	264
Jak to zrobić...	264
Jak to działa...	267
Zobacz również...	267
Co dalej?	268
Wykonywanie połączeń wiele-do-wielu	268
Przygotuj się	268
Jak to zrobić...	269
Jak to działa...	271
Zobacz również...	272
Opracowanie procedury scalania	273
Przygotuj się	273
Jak to zrobić...	273
Jak to działa...	274
Co dalej?	275
Rozdział 9. Porządkowanie i przekształcanie danych	277
Wymagania techniczne	278
Usuwanie zduplikowanych wierszy	278
Przygotuj się...	278
Jak to zrobić...	279
Jak to działa...	281
Zobacz również...	281
Co dalej?	281
Naprawianie relacji wiele-do-wielu	281
Przygotuj się...	282
Jak to zrobić...	282
Jak to działa...	285
Zobacz również...	286
Co dalej?	287

Wykorzystanie stack i melt do zmiany kształtu danych z szerokiego na długi	287
Przygotuj się...	288
Jak to zrobić...	288
Jak to działa...	291
Obracanie wielu grup kolumn	291
Przygotuj się...	291
Jak to zrobić...	292
Jak to działa...	293
Zobacz również...	293
Wykorzystanie unstack i pivot do zmiany kształtu danych z długich na szerokie	294
Przygotuj się...	294
Jak to zrobić...	294
Jak to działa...	296
Rozdział 10. Zdefiniowane przez użytkownika funkcje i klasy do automatyzacji procesu oczyszczania danych	297
Wymagania techniczne	298
Funkcje ułatwiające pierwsze spojrzenie na dane	298
Przygotuj się...	298
Jak to zrobić...	299
Jak to działa...	302
Zobacz również...	302
Funkcje do wyświetlania statystyk podsumowujących i częstości	302
Przygotuj się	303
Jak to zrobić...	303
Jak to działa...	307
Zobacz również...	307
Co dalej?	307
Funkcje do identyfikowania wartości odstających i nieoczekiwanych	308
Przygotuj się	308
Jak to zrobić...	308
Jak to działa...	312
Zobacz również...	313
Co dalej?	313
Funkcje do agregacji lub łączenia danych	313
Przygotuj się	314
Jak to zrobić...	314
Jak to działa...	318
Zobacz również...	318
Co dalej?	318
Klasy zawierające logikę do aktualizowania wartości serii	319
Przygotuj się	319
Jak to zrobić...	319
Jak to działa...	322
Zobacz również...	323
Co dalej?	323
Klasy obsługujące inne niż tabelaryczne struktury danych	324
Przygotuj się	324
Jak to zrobić...	325
Jak to działa...	328
Zobacz również...	328