

Spis treści

O autorze	11
O korektorach merytorycznych	11
Podziękowania	12
Wstęp	13
Co ja tutaj robię?	13
Praktyczna definicja analizy danych	14
Chwila, chwila. A co z big data?	15
Kim jestem?	16
Kim jesteś?	16
Na szczęście będziesz pracować tylko w arkuszu kalkulacyjnym	17
Ale arkusze kalkulacyjne są takie staromodne!	18
Korzystaj z programu Excel lub pakietu LibreOffice	18
Konwencje typograficzne przyjęte w tej książce	19
Zaczynamy	20
1. Wszystko, co chciałeś wiedzieć o arkuszu kalkulacyjnym, ale bałeś się o to zapytać	21
Przykładowe proste dane	22
Szybkie przeglądanie arkusza i klawisz Ctrl	23
Szybkie kopiowanie danych i formuł	24
Formatowanie komórek	26
Wklejanie wartości specjalnych	27
Wstawianie wykresów	28
Menu Znajdź i menu Zamień	29
Formuły przeznaczone do wyszukiwania i wyciągania wartości	30
Stosowanie formuły WYSZUKAJ.PIONOWO do łączenia danych	32
Filtrowanie i sortowanie	33
Stosowanie tabel przestawnych	36
Korzystanie z formuł tablicowych	39
Rozwiązywanie problemów za pomocą narzędzia Solver	40
OpenSolver — chciałbym, abyśmy go nie potrzebowali, ale...	46
Podsumowanie	47
2. Analiza skupień. Część I — zastosowanie algorytmu centroidów do segmentowania bazy klientów	49
Dziewczyny tańczą z dziewczynami, a chłopcy drapią się po łokciach	51
Prawdziwy problem: implementacja algorytmu centroidów w e-mail marketingu	56

Handel winem	56
Początkowy zbiór danych	57
Określanie tego, co chcemy mierzyć	57
Zacznij od czterech grup	61
Odległość euklidesowa — pomiar odległości w linii prostej	61
Odległość dla wszystkich!	64
Określanie położenia środków klastrów	66
Analiza uzyskanych wyników	68
Ustalanie najlepszej oferty dla danego klastra	69
Sylwetka podziału — dobry sposób na określenie optymalnej liczby klastrów	74
A może potrzebujesz pięciu klastrów?	81
Dzielenie klientów na pięć klastrów za pomocą narzędzia Solver	81
Ustalanie najlepszych ofert dla wszystkich pięciu klastrów	82
Określanie sylwetki podziału na pięć klastrów	85
Podział na grupy za pomocą algorytmu k-medioidów i asymetryczny pomiar odległości	87
Podział na grupy za pomocą metody k-medioidów	87
Stosowanie lepszego sposobu pomiaru odległości	87
Implementacja za pomocą Excela	90
Najlepsze oferty przy podziale na pięć klastrów za pomocą median	92
Podsumowanie	95

3. Naiwny klasyfikator bayesowski i niezwykła lekkość bycia idiotą **97**

Jeżeli nazwiesz swój produkt Mandrill, to uzyskasz zaszumione informacje zwrotne	97
Najszybsze na świecie wprowadzenie do rachunku prawdopodobieństwa	100
Obliczanie prawdopodobieństwa warunkowego	100
Prawdopodobieństwo części wspólnej, reguła łańcuchowa i niezależność	101
A co, jeżeli sytuacje są zależne od siebie?	102
Twierdzenie Bayesa	102
Tworzenie modelu sztucznej inteligencji za pomocą twierdzenia Bayesa	103
Zwykle zakłada się, że wysokopoziomowe prawdopodobieństwa klas są sobie równe	105
Kilka innych drobnostek	106
Czas rozpocząć zabawę z Excelem	107
Usuwanie nieistotnych znaków interpunkcyjnych	108
Dzielenie na znakach spacji	109
Zliczanie leksemów i obliczanie prawdopodobieństw	112
Zbudowaliśmy model. Skorzystajmy z niego!	114
Podsumowanie	120

4. Modelowanie optymalizacyjne — „świeżo wyciśnięty” sok nie zamieszka się sam **123**

Dlaczego analityk danych powinien wiedzieć, czym jest optymalizacja?	124
Zacznijmy od prostego kompromisu	125
Przedstawienie problemu w formie wielokomórki	126

Rozwiązywanie problemu poprzez przesuwanie poziomic	128
Metoda simpleks — kręcenie się wokół rogów	129
Praca w Excelu	130
Na końcu tego rozdziału kryje się potwór	140
Szklanka świeżego soku pomarańczowego prosto z drzewa... z przystankiem na modelowanie	141
Trzeba skorzystać z modelu mieszania	142
Zacznijmy od specyfikacji soków	142
Stażność produktu wyjściowego	144
Wprowadzanie danych do Excela	145
Określanie problemu w dodatku Solver	148
Obniżanie standardów	150
Usuwanie cuchnącego problemu — minimalizacja maksymalnych odchyień	154
Warunki i ograniczenie „wielkiego M”	156
Mnożenie zmiennych — skorzystajmy ze 110% mocy Excela	160
Modelowanie ryzyka	168
Dane pochodzące z rozkładu normalnego	168
Podsumowanie	176
5. Analiza skupień. Część II — grafy i analiza sieci	179
Czym jest graf sieci?	180
Wizualizacja prostego grafu	181
Krótkie wprowadzenie do Gephi	184
Instalacja Gephi i przygotowanie pliku	184
Budowa grafu	185
Stopień rozgałęzienia	188
Elegancki wydruk	190
Edycja danych grafu	192
Tworzenie grafu na podstawie danych sprzedaży wina	193
Tworzenie macierzy podobieństwa kosinusowego	195
Generowanie grafu r-sąsiedztwa	197
Jaka jest wartość krawędzi? Nagradzanie i karanie krawędzi — modularność grafu	202
Czym jest punkt, a czym kara?	202
Tworzenie arkusza punktacji	206
Czas dokonać podziału na grupy	208
Podział 1.	208
Podział 2. — kontratak	214
Podział 3. — zemsta	215
Grupy — kodowanie i analiza	216
Tam i z powrotem — czas na Gephi	220
Podsumowanie	225
6. Regresja jako przodek nadzorowanego uczenia maszynowego i sztucznej inteligencji	227
Co? Jesteś w ciąży?	227
Nie oszukuj siebie	228

Przewidywanie ciąży klientów na podstawie regresji liniowej	229
Zbiór cech	230
Tworzenie treningowego zbioru danych	231
Tworzenie zmiennych fikcyjnych	233
Pobawmy się regresją liniową	235
Parametry regresji liniowej: współczynnik determinacji, test F i test t	244
Przewidywanie ciąży na nowym zbiorze danych i sprawdzanie jakości modelu	255
Przewidywanie ciąży klientów za pomocą regresji logistycznej	265
Najpierw musisz określić funkcję wiążącą	265
Tworzenie funkcji logistycznej i ponowna optymalizacja	266
Praca nad prawdziwą regresją logistyczną	270
Wybór modelu — porównywanie skuteczności regresji liniowej i regresji logistycznej	272
Dalsza lektura	274
Podsumowanie	275
7. Modele zespołowe — dużo nie najlepszej pizzy	277
Korzystanie z danych z rozdziału 6.	278
Agregacja — losuj, trenuj, powtórz	280
Pieniek decyzyjny to niezbyt ładne określenie głupiego modelu	280
To wcale nie wydaje się takie głupie!	281
Więcej mocy!	283
Czas rozpocząć proces trenowania	284
Ocena działania modelu zespolonego	293
Wzmacnianie — jeżeli uzyskałeś niesatysfakcjonujące wyniki, to wzmocnij swój model i uruchom go jeszcze raz	298
Trenowanie modelu — każda cecha ma swoje pięć minut	299
Wydajność modelu wzmacnianych reguł decyzyjnych	307
Podsumowanie	311
8. Prognozowanie — oddychaj spokojnie, i tak nie wygrasz	313
Hossa na rynku sprzedaży mieczy	314
Szeregi czasowe	315
Zacznij od prostego wygładzania wykładniczego	317
Przygotowanie arkusza prognozy prostego wygładzania wykładniczego	319
Być może dane zawierają trend	325
Podwójne wygładzanie wykładnicze (metoda Holta)	327
Metoda Holta w arkuszu kalkulacyjnym	329
To wszystko? Analiza autokorelacji	335
Wielokrotne wygładzanie wykładnicze — model Holta-Wintersa	342
Określanie początkowych wartości poziomu, trendu i sezonowości	345
Tworzenie prognozy	349
Czas na optymalizację	354
Powiedz mi, że to już koniec. Proooszę!	356
Interwały prognozy	356
Tworzenie wykresu warstwowego wachlarza wartości	360
Podsumowanie	362

9. Wykrywanie obserwacji odstających — to, że jakiś element jest inny od pozostałych, nie oznacza, że jest nieistotny	365
Element odstający to też człowiek	366
Fascynująca sprawa Hadlumów	367
Metoda Tukeya	368
Implementacja metody Tukeya w arkuszu kalkulacyjnym	368
Ograniczenia tej prostej techniki	371
Nie tragiczny, ale słaby we wszystkim	372
Przygotowywanie danych do utworzenia wykresu	373
Tworzenie grafu	376
Określanie k najbliższych sąsiadów	378
Pierwsza metoda wykrywania elementów odstających grafu — skorzystaj ze stopnia wchodzącego	379
Druga metoda wykrywania elementów odstających grafu — zgłębianie niuansów za pomocą k-odległości	383
Trzecia metoda wykrywania elementów odstających grafu — lokalny miernik stopnia oddalenia obserwacji	385
Podsumowanie	391
10. Przejście z arkusza kalkulacyjnego do języka R	393
Przygotowanie środowiska i początek pracy w języku R	394
Wprowadzanie prostych danych	395
Wczytywanie danych do R	402
Prawdziwa analiza danych	404
Sferyczny algorytm k-średnich wywołany za pomocą zaledwie kilku linii kodu	404
Budowanie modeli sztucznej inteligencji na podstawie danych zakupów (wykrywanie ciąży)	410
Prognozowanie w R	417
Wykrywanie elementów odstających	421
Podsumowanie	426
Wnioski	427
Gdzie ja jestem? Co się stało?	427
Zanim odłożysz tę książkę	428
Poznaj problem	428
Potrzebujemy więcej tłumaczy	429
Uważaj na trójgłowe monstrum: narzędzia, wydajność i perfekcjonizm	430
Nie jesteś najważniejszą osobą w firmie	432
Bądź kreatywny	433
Skorowidz	435