



Spis treści

<i>O autorze</i>	11
<i>O korektorach merytorycznych</i>	11
<i>Podziękowania</i>	12
Wstęp	13
<i>Co ja tutaj robię?</i>	13
<i>Praktyczna definicja analizy danych</i>	14
<i>Chwila, chwila. A co z big data?</i>	15
<i>Kim jestem?</i>	16
<i>Kim jesteś?</i>	16
<i>Na szczęście będziesz pracować tylko w arkuszu kalkulacyjnym</i>	17
<i>Ale arkusze kalkulacyjne są takie staromodne!</i>	18
<i>Korzystaj z programu Excel lub pakietu LibreOffice</i>	18
<i>Konwencje typograficzne przyjęte w tej książce</i>	19
<i>Zaczynamy</i>	20
1. Wszystko, co chciałeś wiedzieć o arkuszu kalkulacyjnym, ale bałeś się o to zapytać	21
<i>Przykładowe proste dane</i>	22
<i>Szybkie przeglądanie arkusza i klawisz Ctrl</i>	23
<i>Szybkie kopiowanie danych i formuł</i>	24
<i>Formatowanie komórek</i>	26
<i>Wklejanie wartości specjalnych</i>	27
<i>Wstawianie wykresów</i>	28
<i>Menu Znajdź i menu Zamień</i>	29

<i>Formuły przeznaczone do wyszukiwania i wyciągania wartości</i>	30
<i>Stosowanie formuły WYSZUKAJ.PIONOWO do łączenia danych</i>	32
<i>Filtrowanie i sortowanie</i>	33
<i>Stosowanie tabel przestawnych</i>	36
<i>Korzystanie z formuł tablicowych</i>	39
<i>Rozwiązywanie problemów za pomocą narzędzia Solver</i>	40
<i>OpenSolver — chciałbym, abyśmy go nie potrzebowali, ale...</i>	46
<i>Podsumowanie</i>	47
2. Analiza skupień. Część I — zastosowanie algorytmu centroidów do segmentowania bazy klientów	49
<i>Dziewczyny tańczą z dziewczynami, a chłopcy drapią się po łokciach</i>	51
<i>Prawdziwy problem: implementacja algorytmu centroidów w e-mail marketingu</i>	56
<i>Handel winem</i>	56
<i>Początkowy zbiór danych</i>	57
<i>Określanie tego, co chcemy mierzyć</i>	57
<i>Zacznij od czterech grup</i>	61
<i>Odległość euklidesowa — pomiar odległości w linii prostej</i>	61
<i>Odległość dla wszystkich!</i>	64
<i>Określanie położenia środków klastrów</i>	66
<i>Analiza uzyskanych wyników</i>	68
<i>Ustalanie najlepszej oferty dla danego klastra</i>	69
<i>Sylwetka podziału — dobry sposób na określenie optymalnej liczby klastrów</i>	74
<i>A może potrzebujesz pięciu klastrów?</i>	81
<i>Dzielenie klientów na pięć klastrów za pomocą narzędzia Solver</i>	81
<i>Ustalanie najlepszych ofert dla wszystkich pięciu klastrów</i>	82
<i>Określanie sylwetki podziału na pięć klastrów</i>	85
<i>Podział na grupy za pomocą algorytmu k-mediodów i asymetryczny pomiar odległości</i>	87
<i>Podział na grupy za pomocą metody k-mediodów</i>	87
<i>Stosowanie lepszego sposobu pomiaru odległości</i>	87
<i>Implementacja za pomocą Excela</i>	90
<i>Najlepsze oferty przy podziale na pięć klastrów za pomocą median</i>	92
<i>Podsumowanie</i>	95
3. Naiwny klasyfikator bayesowski i niezwykła lekkość bycia idiotą	97
<i>Jeżeli nazwiesz swój produkt Mandrill, to uzyskasz zaszumione informacje zwrotne</i>	97
<i>Najszybsze na świecie wprowadzenie do rachunku prawdopodobieństwa</i>	100
<i>Obliczanie prawdopodobieństwa warunkowego</i>	100
<i>Prawdopodobieństwo części wspólnej, reguła łańcuchowa i niezależność</i>	101
<i>A co, jeżeli sytuacje są zależne od siebie?</i>	102
<i>Twierdzenie Bayesa</i>	102

<i>Tworzenie modelu sztucznej inteligencji za pomocą twierdzenia Bayesa</i>	103
<i>Zwykle zakłada się, że wysokopoziomowe prawdopodobieństwa klas są sobie równe</i>	105
<i>Kilka innych drobnostek</i>	106
<i>Czas rozpocząć zabawę z Excelem</i>	107
<i>Usuwanie nieistotnych znaków interpunkcyjnych</i>	108
<i>Dzielenie na znakach spacji</i>	109
<i>Zliczanie leksemów i obliczanie prawdopodobieństw</i>	112
<i>Zbudowaliśmy model. Skorzystajmy z niego!</i>	114
<i>Podsumowanie</i>	120
4. Modelowanie optymalizacyjne — „świeżo wyciśnięty” sok nie zamiesza się sam	123
<i>Dlaczego analityk danych powinien wiedzieć, czym jest optymalizacja?</i>	124
<i>Zacznijmy od prostego kompromisu</i>	125
<i>Przedstawienie problemu w formie wielokomórki</i>	126
<i>Rozwiązywanie problemu poprzez przesuwanie poziomicy</i>	128
<i>Metoda simpleks — kręcenie się wokół rogów</i>	129
<i>Praca w Excelu</i>	130
<i>Na końcu tego rozdziału kryje się potwór</i>	140
<i>Szklanka świeżego soku pomarańczowego prosto z drzewa... z przystankiem na modelowanie</i>	141
<i>Trzeba skorzystać z modelu mieszania</i>	142
<i>Zacznijmy od specyfikacji soków</i>	142
<i>Stołość produktu wyjściowego</i>	144
<i>Wprowadzanie danych do Excela</i>	145
<i>Określanie problemu w dodatku Solver</i>	148
<i>Obniżanie standardów</i>	150
<i>Usuwanie cuchnącego problemu — minimalizacja maksymalnych odchyień</i>	154
<i>Warunki i ograniczenie „wielkiego M”</i>	156
<i>Mnożenie zmiennych — skorzystajmy ze 110% mocy Excela</i>	160
<i>Modelowanie ryzyka</i>	168
<i>Dane pochodzące z rozkładu normalnego</i>	168
<i>Podsumowanie</i>	176
5. Analiza skupień. Część II — grafy i analiza sieci	179
<i>Czym jest graf sieci?</i>	180
<i>Wizualizacja prostego grafu</i>	181
<i>Krótkie wprowadzenie do Gephi</i>	184
<i>Instalacja Gephi i przygotowanie pliku</i>	184
<i>Budowa grafu</i>	185
<i>Stopień rozgałęzienia</i>	188
<i>Elegancki wydruk</i>	190
<i>Edycja danych grafu</i>	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 zespołowego	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
Zaczynj 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
<i>Przygotowanie środowiska i początek pracy w języku R</i>	394
<i>Wprowadzanie prostych danych</i>	395
<i>Wczytywanie danych do R</i>	402
<i>Prawdziwa analiza danych</i>	404
<i>Sferyczny algorytm k-średnich wywołany za pomocą zaledwie kilku linii kodu</i>	404
<i>Budowanie modeli sztucznej inteligencji na podstawie danych zakupów</i> <i>(wykrywanie cięży)</i>	410
<i>Prognozowanie w R</i>	417
<i>Wykrywanie elementów odstających</i>	421
<i>Podsumowanie</i>	426
Wnioski	427
<i>Gdzie ja jestem? Co się stało?</i>	427
<i>Zanim odłożysz tę książkę</i>	428
<i>Poznaj problem</i>	428
<i>Potrzebujemy więcej tłumaczy</i>	429
<i>Uważaj na trójgłowe monstrum: narzędzia, wydajność i perfekcjonizm</i>	430
<i>Nie jesteś najważniejszą osobą w firmie</i>	432
<i>Bądź kreatywny</i>	433
<i>Skorowidz</i>	435