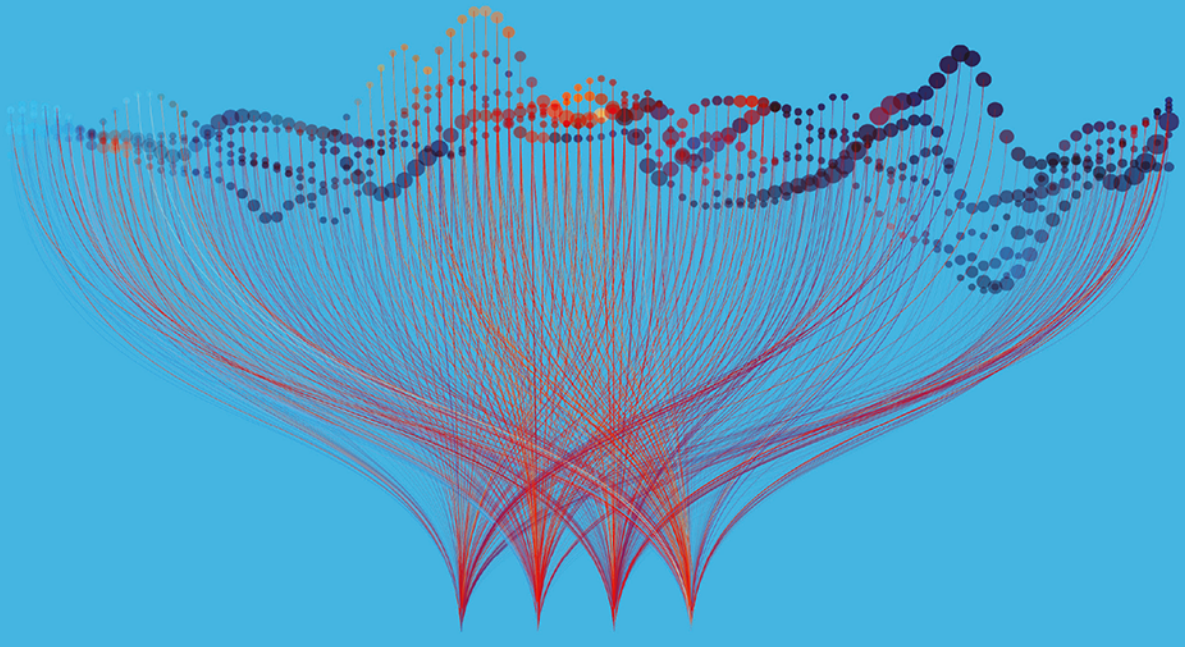
Helion

DATA SCIENCE

Programowanie, analiza i wizualizacja danych z wykorzystaniem języka R



MICHAEL FREEMAN | JOEL ROSS

Tytuł oryginału: Programming Skills for Data Science: Start Writing Code to Wrangle, Analyze, and Visualize Data with R

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-283-5782-2

Authorized translation from the English language edition, entitled PROGRAMMING SKILLS FOR DATA SCIENCE: START WRITING CODE TO WRANGLE, ANALYZE, AND VISUALIZE DATA WITH R, 1st Edition by FREEMAN, MICHAEL; ROSS, JOEL, published by Pearson Education, Inc, publishing as Addison-Wesley Professional, Copyright © 2019 Pearson Education, Inc.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc. POLISH language edition published by HELION SA, Copyright © 2020.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiejkolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz Helion SA dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz Helion SA nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Helion SA
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 231 22 19, 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *http://helion.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
http://helion.pl/user/opinie/datasi
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
ftp://ftp.helion.pl/przyklady/datasi.zip

Printed in Poland.

• Kup książkę
• Poleć książkę
• Oceń książkę

• Księgarnia internetowa
• Lubię to! » Nasza społeczność

Spis treści

Kup książkę Poleć książkę

Kup książkę Poleć książkę

Kup książkę Poleć książkę

Kup książkę Poleć książkę

Kup książkę Poleć książkę

Kup książkę Poleć książkę

9

Jak zrozumieć dane?

We wcześniejszych rozdziałach opisano podstawy programowania na potrzeby pracy z danymi. Szczegółowo wyjaśniono, jak nakazać komputerowi przetwarzanie danych. Aby posłużyć się komputerem do analizy danych, potrzeba *uzyskać dostęp* do zbioru danych i *zinterpretować* je, aby móc zadawać na ich temat sensowne pytania. Dzięki temu możesz przekształcić surowe dane w informacje umożliwiające podejmowanie działań (ang. *actionable information*).

Ten rozdział zawiera ogólny przegląd pokazujący, jak interpretować zbiory danych w kontekście nauki o danych. Szczegółowo opisano tu źródła danych, formaty danych i strategie określania zadawanych pytań. Opracowanie przejrzystego modelu umysłowego znaczenia wartości w zbiorze danych jest niezbędne do zaprogramowania komputera na potrzeby skutecznej analizy danych.

9.1. Proces generowania danych

Zanim zaczniesz pracę z danymi, ważne jest, abyś zrozumiał, *skąd one pochodzą*. Istnieje wiele procesów rejestrowania zdarzeń jako danych. Każdy z tych procesów ma określone ograniczenia i wymaga przyjęcia pewnych założeń. Oto podstawowe kategorie technik rejestrowania danych:

- **Czujniki**. W ostatniej dekadzie ilość danych rejestrowanych przez czujniki gwałtownie wzrosła. Czujniki automatycznie odbierające i rejestrujące informacje, np. czujniki zanieczyszczenia badające jakość powietrza, trafiają do obszaru przetwarzania prywatnych danych (pomyśl o opaskach FitBit lub innych krokomierzach). Jeśli takie urządzenia zostały poprawnie skalibrowane, stanowią niezawodny i spójny mechanizm zbierania danych.
- **Ankiety**. Dane, które trudniej jest mierzyć z zewnątrz, np. opinie lub przeżycia ludzi, można zbierać za pomocą *ankiet*. Ponieważ ankiety są oparte na samoopisie, jakość danych może być różna (w zależności od ankiety i poszczególnych ankietowanych). W niektórych dziedzinach ludzie mogą słabo pamiętać opisywane zagadnienia (np. to, co jedli w zeszłym tygodniu). Czasem mają też motywację do tego, by odpowiadać w konkretny sposób (np. przypisywać sobie więcej prozdrowotnych zachowań, niż rzeczywiście przejawiają). Należy wykrywać tego rodzaju błędy systematyczne i jeśli to możliwe, uwzględniać je w analizach.
- **Przechowywanie rejestrów**. W wielu dziedzinach organizacje stosują automatyczne i ręczne procesy rejestrowania działań. Na przykład szpital może rejestrować długość

Kup książkę Poleć książkę

i skutki każdej przeprowadzanej operacji (a jednostki rządowe mogą wymagać od szpitala raportów na ten temat). Rzetelność takich danych zależy od jakości systemów używanych do ich generowania. Także eksperymenty naukowe wymagają sumiennego rejestrowania wyników.

- **Wtórne analizy danych**. Dane można też kompilować na podstawie *istniejących zbiorów wiedzy* lub pomiarów. Można np. zliczać wystąpienia słów w tekstach historycznych (komputery mogą w tym pomóc!).

Wszystkie te metody zbierania danych mogą prowadzić do problemów i błędów systematycznych. Na przykład czujniki mogą być nieprecyzyjne, ludzie mogą prezentować się w określony sposób w ankietach, rejestry mogą uwzględniać tylko konkretne zadania, a istniejące zbiory wiedzy mogą nie zawierać niektórych punktów widzenia. Gdy pracujesz z jakimś zbiorem danych, koniecznie należy rozważyć, skąd dane pochodzą (*kto* je zarejestrował, *jak* i *dlaczego*), aby móc je skutecznie i sensownie analizować.

9.2. Wyszukiwanie danych

Możliwość rejestrowania i przechowywania danych w komputerach doprowadziła do gwałtownego wzrostu ilości danych, jakie można analizować — od osobistych pomiarów biologicznych (*ile kroków zrobiłem?*), przez struktury sieci społecznościowych (*kim są moi znajomi?*), po prywatne informacje, które wyciekły z niezabezpieczonych witryn i agencji rządowych (*jakie są numery PESEL użytkowników?*). W pracy zapewne będziesz korzystał z wewnętrznych danych zbieranych lub zarządzanych przez Twoją firmę. Mogą to być bardzo różnorodne dane: od zamówień w kawiarni fair trade po wyniki badań medycznych. Takie dane mogą być tak różne jak rodzaje firm, ponieważ obecnie *każdy* rejestruje dane i dostrzega wartość ich analizy.

Na szczęście istnieje jest też mnóstwo bezpłatnych, powszechnie dostępnych zbiorów danych, z którymi można pracować. Organizacje często udostępniają publicznie duże ilości danych, aby umożliwiać powtarzanie eksperymentów, promować przejrzystość lub po prostu sprawdzać, jak inni potrafią wykorzystać określone dane. Takie zbiory danych doskonale nadają się do budowania umiejętności i portfolio w zakresie nauki o danych. Dane te są dostępne w różnych formatach, np. jako arkusze CSV (zobacz rozdział 10.), relacyjne bazy danych (zobacz rozdział 13.) i interfejsy API usług sieciowych (zobacz rozdział 14.).

Oto popularne źródła publicznie dostępnych zbiorów danych:

- **Publikacje rządowe**. Organizacje rządowe (i inne systemy biurokratyczne) w ramach codziennej działalności generują *mnóstwo* danych i często udostępniają je, aby uchodzić za przejrzyste i odpowiedzialne instytucje przed opinią publiczną. Obecnie dane są publicznie dostępne w wielu krajach, na przykład w Stanach Zjednoczonych[1], Kanadzie[2], Indiach[3] itd. Także lokalne instytucje udostępniają dane. Na przykład miasto Seattle[4] udostępnia duże ilości danych w łatwo dostępnym formacie. Dane rządowe dotyczą

[1] **Publiczne dane rządu Stanów Zjednoczonych**: *https://www.data.gov*.

[2] **Publiczne dane rządu Kanady**: *https://open.canada.ca/en/open-data*.

[3] **Platforma publicznych danych rządu Indii**: *https://data.gov.in*.

[4] **Portal publicznych danych miasta Seattle**: *https://data.seattle.gov*.

Kup książkę Poleć książkę

różnych zagadnień, przy czym wpływ na treść danych może mieć sytuacja polityczna związana z ich zbieraniem i przechowywaniem.

- **Wiadomości i dzienniki**. Dziennikarstwo pozostaje jednym z najważniejszych kontekstów, w jakich dane są zbierane i analizowane. Dziennikarze wykonują dużo czarnej roboty w obszarze generowania danych — przeszukują istniejące źródła wiedzy, przepytują ludzi, przeprowadzają ankiety oraz w inny sposób ujawniają i łączą wcześniej ukryte lub ignorowane informacje. Media informacyjne zwykle publikują przeanalizowane, podsumowujące informacje do użytku odbiorców, jednak czasem udostępniają też źródła danych, aby inni mogli potwierdzać i rozwijać wnioski z analiz. Na przykład *New York Times*[5] udostępnia dużą część danych historycznych w usługach sieciowych, a poświęcony danym politycznym blog *FiveThirtyEight*[6] udostępnia w serwisie GitHub wszystkie dane (błędne modele i inne informacje), których dotyczą publikowane na tym blogu artykuły.
- **Badania naukowe**. Innym doskonałym źródłem danych są badania naukowe prowadzone na uczelniach lub przez instytucje komercyjne. Badania naukowe są (teoretycznie) dobrze uzasadnione i ustrukturyzowane oraz zapewniają sensowne dane, jeśli są one analizowane w odpowiednim dla nich kontekście. Ponieważ badania muszą być upowszechniane, a następnie sprawdzane przez inne osoby, aby były użyteczne, często są publicznie udostępniane na potrzeby analiz i krytyki. Niektóre czasopisma naukowe, np. poważany magazyn *Nature*, wymagają, aby autorzy udostępniali dane innym do sprawdzenia (zapoznaj się z listą repozytoriów danych naukowych tego magazynu[7]).
- **Sieci i media społecznościowe**. Internet jest jednym ze źródeł, gdzie generowane są największe ilości danych. Dane są tu rejestrowane automatycznie w wyniku użytkowania aplikacji społecznościowych takich jak Facebook, Twitter i Google oraz interakcji w tych narzędziach. Aby lepiej zintegrować swoje usługi z codziennym życiem użytkowników, firmy zarządzające mediami społecznościowymi udostępniają swoje dane programowo innym programistom. Możesz np. uzyskać dostęp na żywo do danych z Twittera[8], które są używane do różnych ciekawych analiz. Także Google zapewnia programowy dostęp[9] do wielu swoich usług (w tym do wyszukiwarki i serwisu YouTube).
- **Społeczności internetowe**. Wraz z szybkim wzrostem popularności nauki o danych rozwija się też społeczność specjalistów z tej dziedziny. Ta społeczność i używane przez nią serwisy internetowe są następnym doskonałym źródłem ciekawych i różnorodnych zbiorów danych oraz analiz. Na przykład serwis *Kaggle*[10] udostępnia wiele zbiorów

[5] **Serwis dla programistów prowadzony przez dziennik *New York Times***: *https://developer.nytimes.com*.

[6] ***FiveThirtyEight* — Our Data**: *https://data.fivethirtyeight.com*.

[7] **Nature — polecane repozytoria danych**: *https://www.nature.com/sdata/policies/repositories*.

[8] **Platforma dla programistów w serwisie Twitter**: *https://developer.twitter.com/en/docs*.

[9] **Przegląd interfejsów API Google'a**: *https://developers.google.com/apis-explorer/*.

[10] **Kaggle** — the home of data science and machine learning: *https://www.kaggle.com*.

Kup książkę Poleć książkę

danych, a także zadań związanych z ich analizą. Serwis *Socrata*[11] (gdzie znajduje się repozytorium danych miasta Seattle) także rejestruje różnorodne zbiory danych (często od jednostek komercyjnych i rządowych). Nieco podobny serwis *UCI Machine Learning Repository*[12] zawiera zbiory danych używanych do uczenia maszynowego. Pochodzą one głównie ze źródeł akademickich. Istnieje też wiele innych internetowych list źródeł danych, w tym specjalna lista dyskusyjna w serwisie Reddit — `/r/Datasets`[13].

Dostępnych jest więc mnóstwo pochodzących z rzeczywistego świata zbiorów danych, nad którymi możesz pracować — niezależnie od tego, czy szukasz odpowiedzi na konkretne pytania, czy chcesz tylko eksplorować dane lub szukasz pomysłów.

9.3. Rodzaje danych

Gdy masz już zbiór danych, musisz zrozumieć jego strukturę i zawartość, zanim zaczniesz go (programowo) badać. Zrozumienie *rodzajów* danych zależy od umiejętności rozróżniania skal pomiarowych, a także różnych struktur używanych do przechowywania danych.

9.3.1. Skale pomiarowe

Dane mogą obejmować różne rodzaje wartości (reprezentowane w R za pomocą typów danych). Na ogólnym poziomie wartości można omawiać w kategoriach **skal pomiarowych**[14] — sposobu klasyfikowania wartości danych w kategoriach tego, jak można je mierzyć i jak porównywać je z innymi wartościami.

W dziedzinie statystyki wartości są zwykle dzielone na cztery kategorie opisane w tabeli 9.1.

Dane nominalne nie mają określonego uporządkowania. Nie można np. powiedzieć, że „jabłka są bardziej niż pomarańcze", choć można stwierdzić, czy dany owoc jest jabłkiem lub pomarańczą. Dane nominalne często służą do określania, że dana obserwacja należy do konkretnej kategorii lub grupy. Na danych nominalnych zwykle nie przeprowadza się analiz matematycznych (np. nie da się ustalić „średniej" na podstawie owoców). Można jednak analizować liczby wystąpień lub rozkłady. Dane nominalne można reprezentować za pomocą łańcuchów znaków (np. nazw owoców), ale też liczb (np. „pierwszy rodzaj owoców", „drugi rodzaj owoców" itd.). Jednak to, że wartość w zbiorze danych jest liczbą, nie oznacza, że możesz wykonywać operacje matematyczne. Zwróć też uwagę, że wartości logiczne (`TRUE` i `FALSE`) są wartościami nominalnymi.

W **danych porządkowych** określona jest *kolejność* wartości. Takie dane można stosować do kategoryzowania, ale też do określania, że niektóre grupy są *większe* lub *mniejsze* od innych. Na przykład hotele lub restauracje można sklasyfikować jako *pięciogwiazdkowe*, *czterogwiazdkowe* itd. W takich danych występuje uporządkowanie, jednak odległości między wartościami mogą być różne. Możliwe jest tu wyznaczenie wartości minimalnej, maksymalnej, a nawet mediany,

---

[11] **Socrata** — platforma typu „dane jako usługa": *https://opendata.socrata.com*.

[12] **UCI Machine Learning Repository**: *https://archive.ics.uci.edu/ml/index.php*.

[13] **/r/DataSets**: *https://www.reddit.com/r/datasets/*.

[14] Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677 – 680. *https://doi.org/10.1126/science.103.2684.677*.

Kup książkę Poleć książkę

Tabela 9.1. Skale pomiarowe

Skala	Przykład	Operacje
Nominalna Bez kolejności, służy do klasyfikowania.	*Owoce*: jabłka, banany, pomarańcze itd.	==, != „takie same lub różne”
Porządkowa Kolejność, można sortować.	*Oceny hoteli*: pięciogwiazdkowe, czterogwiazdkowe itd.	==, !=, <, > „takie same lub różne”
Ilorazowa (stosunkowa) Kolejność, ustalony punkt zero.	Długości: 1 cal, 1,5 cala, 2 cale itd.	==, !=, <, >, +, -, *, / „dwa razy takie”
Interwałowa (przedziałowa) Kolejność, bez ustalonego punktu zero.	*Daty*: 15/05/2012, 17/04/2015 itd.	==, !=, <, >, +, - „trzy jednostki większe”

jednak nie da się obliczyć średniej statystycznej (ponieważ wartości porządkowe nie definiują, *o ile* jedna wartość jest większa od innej). Zauważ, że zmienne nominalne można traktować jak porządkowe, jeśli wyznaczona zostanie kolejność. Skutkuje to jednak zmianą skali pomiarowej danych. Na przykład kolory tworzą zwykle skalę *nominalną*. Nie można stwierdzić, że „czerwony jest większy od niebieskiego”. Jest tak mimo zwyczajowego porządkowania opartego na kolorach tęczy. Gdy mówisz, że „czerwony występuje przed niebieskim (w tęczy)”, zastępujesz nominalną skalę kolorów skalą porządkową, reprezentującą *pozycje kolorów w tęczy* (które same zależą od *stosunku* długości fal). Dane porządkowe są jednocześnie danymi nominalnymi.

Skala ilorazowa (czasem nazywana **stosunkową**) to najczęściej używana w praktyce skala pomiarowa. Dane dotyczące liczby wystąpień w populacji, wartości pieniężnych lub ilości aktywności są zwykle mierzone z użyciem skali ilorazowej. Dla takich danych można obliczać średnie, a także mierzyć odległości między różnymi wartościami (pomiar odległości jest możliwy także dla danych interwałowych). Zgodnie z oczekiwaniami dane ilorazowe umożliwiają określanie stosunku dwóch wartości (np. wartość x jest dwa razy większa od wartości y). Dane ilorazowe są uznawane za ciągłe.

Skala interwałowa (inaczej **przedziałowa**) jest podobna do ilorazowej, jednak nie występuje tu ustalony punkt zero. Na przykład dat nie można analizować w ujęciu *proporcjonalnym* (nie możesz np. stwierdzić, że *środa* jest dwa razy większa od *poniedziałku*). Możliwe jest obliczenie odległości (interwału) między dwoma wartościami (np. *dwa dni*), jednak nie da się ustalić *stosunku* tych wartości. Dane interwałowe także są uznawane za ciągłe.

Identyfikowanie i rozumienie skal pomiarowych dla określonych atrybutów danych jest ważne, gdy rozważasz, jak analizować zbiór danych. Przede wszystkim musisz wiedzieć, jakiego rodzaju analizy statystyczne będą poprawne dla określonych danych. Musisz też wiedzieć, jak interpretować zjawiska mierzone za pomocą tych danych.

9.3.2. Struktury danych

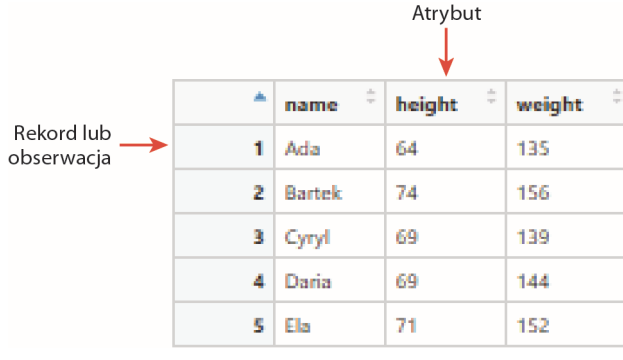
W praktyce będziesz musiał porządkować opisane w poprzednich rozdziałach liczby, łańcuchy znaków, wektory i listy wartości z wykorzystaniem bardziej złożonych formatów. Dane są porządkowane za pomocą pomocnych **struktur** (zwłaszcza gdy zbiór danych jest duży), aby lepiej

Kup książkę Poleć książkę

pokazać, co reprezentują określone liczby i łańcuchy znaków. Na potrzeby pracy z danymi z rzeczywistego świata musisz zrozumieć takie struktury i terminologię używaną do ich opisu.

W praktyce większość zbiorów danych jest przechowywana jako **tabele** z informacjami, a poszczególne wartości danych są uporządkowane w *wiersze* i *kolumny* (zobacz rysunek 9.1). Te tabele przypominają arkusze kalkulacyjne z programów takich jak Microsoft Excel. W tabeli każdy wiersz reprezentuje **rekord** lub **obserwację** — wystąpienie jednej mierzonej rzeczy, np. osoby lub spotkania sportowego. Każda kolumna reprezentuje **atrybut** — określoną właściwość lub aspekt mierzonej rzeczy, np. wagę lub wzrost człowieka albo wynik meczu. Każdą wartość danych można powiązać z **komórką** takiej tabeli.

	name	height	weight
1	Ada	64	135
2	Bartek	74	156
3	Cyryl	69	139
4	Daria	69	144
5	Ela	71	152



Rysunek 9.1. Tabela danych (wagi i wzrostu osób). Wiersze reprezentują obserwacje, natomiast kolumny odpowiadają atrybutom

W tym ujęciu tabela jest kolekcją mierzonych rzeczy. Dla każdej cechy danej rzeczy określona jest konkretna wartość. Ponieważ wszystkie obserwacje mają te same cechy (atrybuty), można analizować je porównawczo. Ponadto dzięki uporządkowaniu danych w tabeli każdej wartości (komórce) można automatycznie nadać dwa znaczenia: z której obserwacji pochodzi i jaki atrybut reprezentuje. Ta struktura umożliwia wydobycie znaczenia semantycznego na podstawie liczb. Liczba 64 na rysunku 9.1 nie jest przypadkową wartością — to „wzrost Ady".

Tabela z rysunku 9.1 reprezentuje *niewielki* (a nawet malutki) zbiór danych, obejmujący tylko pięć obserwacji (wierszy). Wielkość zbioru danych jest zwykle mierzona na podstawie liczby obserwacji. Niewielki zbiór danych obejmuje kilkadziesiąt obserwacji, natomiast w dużym zbiorze mogą znajdować się tysiące lub setki tysięcy rekordów. Pojęcie „big data" oznacza zbiory danych, które są tak duże, że nie można ich wczytać do pamięci komputera bez specjalnych zabiegów. Takie zbiory mogą obejmować miliardy, a nawet tryliony wierszy. Jednak nawet

Kup książkę Poleć książkę

zbiory danych o stosunkowo niewielkiej liczbie obserwacji mogą zawierać dużą liczbę komórek, jeśli dla każdej obserwacji rejestrowanych jest wiele atrybutów (przy czym takie tabele często można „obrócić", aby wierszy było więcej, a kolumn mniej; zobacz rozdział 12.). Liczba obserwacji i atrybutów (wierszy i kolumn) wyznacza **wymiary** zbioru danych. Nie należy mylić tych wymiarów z „dwuwymiarową" strukturą danych tabeli (struktura jest dwuwymiarowa, ponieważ każda wartość oznacza *dwie* rzeczy: obserwację i atrybut).

Choć dane często są strukturyzowane w przedstawiony sposób, nie muszą być reprezentowane za pomocą jednej tabeli. Złożone zbiory danych mogą obejmować dane zapisane w wielu tabelach (np. w bazie danych; zobacz rozdział 13.). W innych złożonych strukturach danych każda komórka tabeli może obejmować wektor, a nawet całą tabelę. Wtedy tabela nie jest już dwuwymiarowa, ale trójwymiarowa lub o większej liczbie wymiarów. Wiele zbiorów danych udostępnianych przez usługi sieciowe ma strukturę zagnieżdżonych tabel. Informacje na ten temat zawiera rozdział 14.

9.4. Interpretowanie danych

Pierwszą rzeczą, jaką należy zrozumieć po pobraniu zbioru danych (który znalazłeś w internecie lub otrzymałeś w firmie), jest zrozumienie znaczenia danych. Wymaga to zrozumienia dziedziny, w jakiej pracujesz, a także konkretnego użytego schematu danych.

9.4.1. Zdobywanie wiedzy w danej dziedzinie

Pierwszy krok w kierunku zrozumienia zbioru danych polega na zbadaniu i poznaniu dziedziny problemu. **Dziedzina problemu** to zestaw zagadnień powiązanych z danym problemem, czyli kontekst danych. Praca z danymi wymaga **wiedzy z zakresu dziedziny**. Musisz choć na podstawowym poziomie rozumieć dziedzinę problemu, aby móc przeprowadzić jakiekolwiek sensowne analizy danych. Musisz też opracować model umysłowy dotyczący znaczenia danych. Należy zrozumieć znaczenie i funkcje wszystkich atrybutów (aby nie wykonywać obliczeń na pozbawionych kontekstu liczbach), zakres oczekiwanych wartości poszczególnych atrybutów (na potrzeby wykrywania wartości odstających i innych błędów), a także subtelności, które mogą nie być od razu widoczne w zbiorze danych (np. błędy systematyczne lub agregacje, które mogą ukrywać ważne związki przyczynowo-skutkowe).

Oto konkretny przykład: jeśli chcesz przeanalizować tabelę z rysunku 9.1, musisz najpierw zrozumieć, co oznacza „wzrost" i „waga" osoby, używane jednostki (cale, centymetry, a może coś innego?), oczekiwany zakres wartości (czy wzrost 64 Ady oznacza, że jest ona niska?) i inne zewnętrzne czynniki, które mogą wpływać na dane (np. wiek osób).

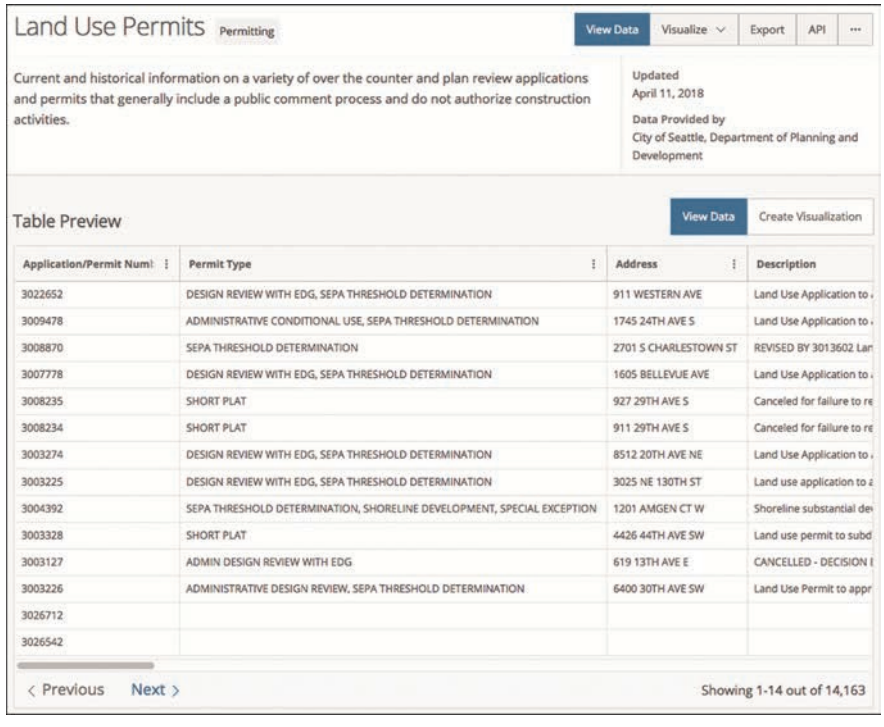
> **Zapamiętaj**: Nie musisz być ekspertem w dziedzinie problemu (choć z pewnością nie zaszkodzi nim być). Musisz jedynie posiadać *wystarczającą* wiedzę, aby pracować z danymi z tej dziedziny.

Choć dane o wzroście ludzi i inne zbiory danych z tej książki powinny być dla większości czytelników zrozumiałe, w praktyce z dużym prawdopodobieństwem natrafisz na dziedziny problemu, w których nie jesteś ekspertem. Jeszcze bardziej problematyczne jest to, że może

Kup książkę Poleć książkę

Ci się *wydawać*, że rozumiesz daną dziedzinę problemu, jednak w rzeczywistości posiadasz jej błędny model umysłowy (jest to błąd w obszarze *metapoznania*).

Rozważ np. zbiór danych z rysunku 9.2 (jest to zrzut tabeli z repozytorium danych miasta Seattle). Ten zbiór danych zawiera informacje na temat zezwoleń na użytkowanie gruntu, związanych z dość skomplikowaną procedurą biurokratyczną, z którą wiele osób może nie być zaznajomionych. Oto pytanie: jak zdobyć wystarczającą wiedzę z dziedziny problemu, aby móc zrozumieć i przeanalizować ten zbiór danych?



Land Use Permits Permitting

View Data | Visualize | Export | API | ...

Current and historical information on a variety of over the counter and plan review applications and permits that generally include a public comment process and do not authorize construction activities.

Updated
April 11, 2018

Data Provided by
City of Seattle, Department of Planning and Development

Table Preview

View Data | Create Visualization

Application/Permit Numb	Permit Type	Address	Description
3022652	DESIGN REVIEW WITH EDG, SEPA THRESHOLD DETERMINATION	911 WESTERN AVE	Land Use Application to
3009478	ADMINISTRATIVE CONDITIONAL USE, SEPA THRESHOLD DETERMINATION	1745 24TH AVE S	Land Use Application to
3008870	SEPA THRESHOLD DETERMINATION	2701 S CHARLESTOWN ST	REVISED BY 3013602 Lan
3007778	DESIGN REVIEW WITH EDG, SEPA THRESHOLD DETERMINATION	1605 BELLEVUE AVE	Land Use Application to
3008235	SHORT PLAT	927 29TH AVE S	Canceled for failure to re
3008234	SHORT PLAT	911 29TH AVE S	Canceled for failure to re
3003274	DESIGN REVIEW WITH EDG, SEPA THRESHOLD DETERMINATION	8512 20TH AVE NE	Land Use Application to
3003225	DESIGN REVIEW WITH EDG, SEPA THRESHOLD DETERMINATION	3025 NE 130TH ST	Land use application to a
3004392	SEPA THRESHOLD DETERMINATION, SHORELINE DEVELOPMENT, SPECIAL EXCEPTION	1201 AMGEN CT W	Shoreline substantial de
3003328	SHORT PLAT	4426 44TH AVE SW	Land use permit to subd
3003127	ADMIN DESIGN REVIEW WITH EDG	619 13TH AVE E	CANCELLED - DECISION
3003226	ADMINISTRATIVE DESIGN REVIEW, SEPA THRESHOLD DETERMINATION	6400 30TH AVE SW	Land Use Permit to appr
3026712			
3026542			

< Previous Next >

Showing 1-14 out of 14,163

Rysunek 9.2. Podgląd danych na temat zezwoleń na użytkowanie gruntu w Seattle[15]. Zawartość została zmodyfikowana na potrzeby tej książki

Zdobywanie wiedzy z dziedziny prawie zawsze wymaga nauki z zewnętrznych źródeł. Rzadko uda Ci się zrozumieć dziedzinę na podstawie samych analiz liczb z arkusza kalkulacyjnego. Aby zdobyć ogólną wiedzę z dziedziny, warto zacząć od ogólnego źródła wiedzy, *Wikipedii*, gdzie łatwo znajdziesz podstawowe opisy. Koniecznie przeczytaj powiązane artykuły lub materiały, aby lepiej zrozumieć zagadnienie. Aby przejrzeć obszerne źródła informacji w internecie, należy zapoznać się z różnymi materiałami, do których prowadzą odsyłacze, i przełożyć znalezione informacje na używany zbiór danych.

[15] **Miasto Seattle — zezwolenia na użytkowanie gruntu** (aby uzyskać dostęp, trzeba założyć bezpłatne konto): *https://data.seattle.gov/Permitting/Land-Use-Permits/uyyd-8gak*.

Kup książkę Poleć książkę

Jednak najlepszym sposobem na poznanie dziedziny problemu jest znalezienie *eksperta z dziedziny*, który objaśni Ci dane zagadnienie. Jeśli chcesz dowiedzieć się czegoś o pozwoleniach na użytkowanie gruntu, spróbuj znaleźć kogoś, kto korzystał z takiego pozwolenia. Drugie w kolejności rozwiązanie polega na zapytaniu się bibliotekarza. Bibliotekarze uczą się pomagać ludziom w znajdowaniu i zdobywaniu podstawowej wiedzy z dziedziny. W bibliotece możesz też znaleźć wyspecjalizowane źródła informacji.

9.4.2. Jak zrozumieć schematy danych?

Gdy już na ogólnym poziomie rozumiesz kontekst dla zbioru danych, możesz zacząć je interpretować. Musisz skupić się na zrozumieniu **schematu danych** (czyli tego, co reprezentują wiersze i kolumny), a także kontekstu wartości. Zachęcamy, aby w analizach zadawać sobie następujące pytania:

Jakie metadane *są dostępne dla zbioru danych?*

Wiele publicznie dostępnych zbiorów danych jest powiązanych z podsumowującymi objaśnieniami, instrukcjami dostępu i użytkowania, a nawet opisami poszczególnych atrybutów. Tego rodzaju **metadane** (czyli dane na temat danych) pochodzą bezpośrednio od źródła danych, dlatego są najlepszym sposobem na zrozumienie, jakie wartości są reprezentowane w poszczególnych komórkach tabeli.

Na przykład na stronie dotyczącej zezwoleń na użytkowanie gruntu w Seattle znajdują się: krótkie podsumowanie (choć warto sprawdzić, co oznacza punkt „over-the-counter review application"), liczba kategorii i oznaczeń, wymiary zbioru danych (w czasie powstawania tej książki dostępnych jest 14 200 wierszy) i krótkie opisy wszystkich kolumn.

Wyjątkowo ważnym aspektem metadanych, który należy sprawdzić, jest następująca kwestia:

Kto utworzył ten zbiór danych? Skąd pochodzą dane?

Zrozumienie tego, kto wygenerował zbiór danych (i jak to zrobił) pozwoli Ci ustalić, gdzie znaleźć więcej informacji na temat tych danych. Dowiesz się też, kim są eksperci z określonej dziedziny. Znajomość źródła danych i metod ich zbierania może też pomóc Ci odkryć ukryte błędy systematyczne lub inne subtelności, które nie są oczywiste na podstawie samych danych. Na przykład na stronie dotyczącej zezwoleń na użytkowanie gruntu napisano, że dane zostały udostępnione przez dział planowania i rozwoju miasta Seattle (obecnie dział budownictwa i inspekcji — Department of Construction & Inspections). Jeśli poszukasz tej jednostki, może znajdziesz jej witrynę[16]. Jest ona dobrym miejscem do tego, by znaleźć więcej informacji na temat danych z używanego zbioru.

Gdy już zrozumiesz metadane, możesz zacząć analizować same dane:

Jakie atrybuty znajdują się w zbiorze danych?

[16] **Dział budownictwa i inspekcji miasta Seattle** (dostęp wymaga założenia bezpłatnego konta): *http://www.seattle.gov/dpd/*.

Kup książkę Poleć książkę

Niezależnie od dostępności metadanych musisz zrozumieć kolumny tabeli. Przyjrzyj się każdej kolumnie i sprawdź, czy znasz odpowiedzi na następujące pytania:

1. Jaki atrybut z rzeczywistego świata dana kolumna ma reprezentować?
2. W jakich jednostkach podawane są wartości (dotyczy to danych ciągłych)?
3. Jakie kategorie są reprezentowane i co one oznaczają (dotyczy to danych nominalnych)?
4. Jaki jest możliwy zakres wartości?

Jeśli metadane zawierają *legendę* tabeli, zadanie jest łatwe. W przeciwnym razie do zrozumienia atrybutów konieczne może być przeanalizowanie źródła danych, co wymaga dodatkowych badań nad dziedziną problemu.

Wskazówka: Gdy przeglądasz zbiór danych — a tak naprawdę niemal dowolne materiały — powinieneś *zapisywać* pojęcia i wyrażenia, których nie znasz, aby móc je później sprawdzić. Dzięki temu nie będziesz (nieprawidłowo) domyślał się znaczenia pojęć i łatwiej rozdzielisz nazwy, które już znasz, od tych jeszcze nieznanych.

Na przykład w metadanych w zbiorze danych dotyczącym zezwoleń na użytkowanie gruntu znajdują się jednoznaczne opisy kolumn, jednak gdy przyjrzysz się przykładowym danym, zobaczysz, że niektóre wartości mogą wymagać dodatkowych analiz. Na przykład czym są różne rodzaje zezwoleń (Permit Types) i rodzaje decyzji (Decision Types)? Gdy wrócisz do źródła danych (na stronę główną działu budownictwa), możesz przejść do strony zezwoleń (Permits), a następnie kliknąć Permits We Issue (A-Z), aby zobaczyć pełną listę rodzajów zezwoleń. Dzięki temu dowiesz się np., że zezwolenie PLAT dotyczy „tworzenia lub modyfikowania poszczególnych działek w ramach nieruchomości", czyli zmiany granicy działek.

Aby zrozumieć atrybuty, musisz przyjrzeć się przykładowym obserwacjom. Otwórz arkusz kalkulacyjny lub tabelę i przejrzyj kilka pierwszych wierszy, aby zobaczyć, jakiego rodzaju wartości występują i co może z nich wynikać.

W całym procesie nieustannie powinieneś zadawać sobie pytanie:

Jakich pojęć nie znam lub nie rozumiem?

W zależności od dziedziny problemu zbiór danych może obejmować dużą ilość *żargonu* — zarówno w samych danych, jak i w ich objaśnieniach. Upewnij się, że rozumiesz wszystkie używane pojęcia techniczne. Jest to bardzo ważna kwestia, jeśli chcesz skutecznie omawiać i analizować dane.

Ostrzeżenie: Zwracaj uwagę na akronimy, których nie znasz, i koniecznie sprawdzaj ich znaczenie.

Na przykład w podglądzie tabeli (Table Preview) wiele wartości atrybutu Permit Type to SEPA. Gdy poszukasz tego akronimu, znajdziesz stronę z opisem ustawy *State Policy Environmental Act* (wymaga ona uwzględnienia wpływu sposobu użytkowania gruntu na środowisko), a także informacje na temat procesu „określania wartości progowej" (Threshold Determination).

Interpretowanie zbioru danych wymaga więc analiz i pracy, która *nie* jest programowaniem. Choć może się wydawać, że taka praca utrudnia Ci postępy w przetwarzaniu danych, poprawny model umysłowy danych jest przydatny i niezbędny do przeprowadzania analiz.

Kup książkę Poleć książkę

9.5. Odpowiadanie na pytania na podstawie danych

Prawdopodobnie najtrudniejszym aspektem analizy danych jest skuteczne zadawanie pytań na podstawie danych, aby uzyskać pożądane informacje. Specjalista od nauki o danych często odpowiada za przekładanie różnych pytań z dziedziny na konkretne obserwacje i atrybuty w zbiorze danych. Przyjrzyj się np. następującemu pytaniu:

Jaka jest najgorsza choroba w Stanach Zjednoczonych?

Aby odpowiedzieć na to pytanie, trzeba zrozumieć dziedzinę problemu — pomiary obciążenia chorobami — i znaleźć zbiór danych, który nadaje się do udzielenia odpowiedzi. Odpowiednim zbiorem danych może być tu badanie *Global Burden of Disease*[17] przeprowadzone przez jednostkę Institute for Health Metrics and Evaluation. Szczegółowo opisano tam *obciążenie chorobami* w Stanach Zjednoczonych i na świecie.

Po pobraniu tego zbioru danych musisz **zoperacjonalizować** podstawowe pytanie. Przeanalizujmy każde kluczowe słowo — musisz zidentyfikować zbiór *chorób*, a następnie zdefiniować, co oznacza, że choroba jest „najgorsza". Pytanie można ująć bardziej konkretnie, co ilustrują poniższe interpretacje:

- Która choroba powoduje *największą liczbę zgonów* w Stanach Zjednoczonych?
- Która choroba powoduje *największą liczbę przedwczesnych zgonów* w Stanach Zjednoczonych?
- Która choroba powoduje *największą liczbę niepełnosprawności* w Stanach Zjednoczonych?

W zależności od definicji „najgorszej" choroby będziesz wykonywał zupełnie inne obliczenia i analizy. Zapewne uzyskasz też inne odpowiedzi. Dlatego musisz umieć zdecydować, *co dokładnie oznacza dane pytanie*. Zadanie to wymaga zrozumienia niuansów z dziedziny problemu związanej z pytaniem.

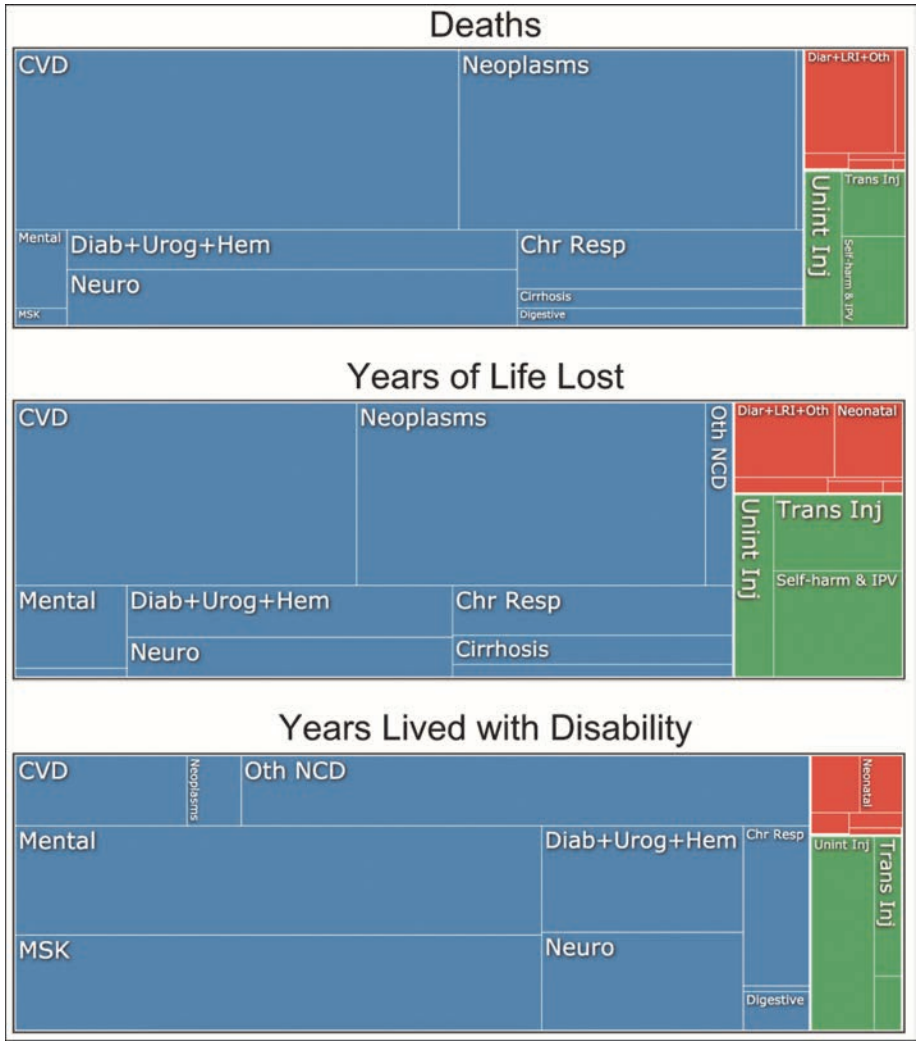
Na rysunku 9.3 pokazano wizualizacje, które pomagają odpowiedzieć na zadane pytanie. Na rysunku znajdują się zrzuty *map drzewa* (ang. *treemap*) z internetowego narzędzia *GBD Compare*[18]. Mapa drzewa przypomina wykres kołowy, ale jest zbudowana z prostokątów. Powierzchnia każdego segmentu jest proporcjonalna do wartości danych. Dodatkową zaletą map drzew jest to, że pozwalają pokazywać *hierarchie* informacji dzięki *zagnieżdżaniu* różnych poziomów prostokątów jedne w drugich. Na przykład na rysunku 9.3 obciążenie chorobami zakaźnymi (wyróżnionymi na czerwono) jest przedstawione w jednym segmencie wykresu.

W zależności od tego, jak zoperacjonalizujesz „najgorszą chorobę", wybrana zostanie inna z nich. Na rysunku 9.3 widać, że prawie 90% wszystkich zgonów jest spowodowanych chorobami niezakaźnymi (kolor niebieski), takimi jak choroby układu krążenia (*CVD*) i nowotwory (*Neoplasms*). Gdy uwzględnić wiek pacjentów (za pomocą wskaźnika *utracone lata życia*), odsetek zgonów powodowanych przez te choroby spada do 80%. Ten wskaźnik pozwala też zidentyfikować przyczyny zgonów, które w nieproporcjonalny sposób wpływają na młodych ludzi. Te przyczyny to np. wypadki drogowe (*Trans Inj*) i samookaleczenia. Te czynniki są

[17] **IHME — globalne obciążenie chorobami**: *http://www.healthdata.org/node/835*.

[18] **GBD Compare** — wizualizacja globalnego obciążenia chorobami: *https://vizhub.healthdata.org/gbd-compare/*.

Kup książkę Poleć książkę



Rysunek 9.3. Mapy drzewa z narzędzia GBD Compare. Widoczne są tu proporcje zgonów (u góry), utraconych lat życia (środek) i lat przeżytych z niepełnosprawnością (u dołu) powodowanych przez różne choroby w Stanach Zjednoczonych

wyróżnione kolorem zielonym (zobacz środkowy wykres na rysunku 9.3). Z kolei jeśli za „najgorsze" choroby uznać te, które powodują najwięcej fizycznych niepełnosprawności w populacji (dolny wykres na rysunku 9.3), ujawnia się wpływ chorób układu mięśniowo-kostnego (*MSK*) i zaburzeń umysłowych (*Mental*).

Ponieważ analizy danych służą do znajdowania odpowiedzi na pytania, pierwszy krok polega na upewnieniu się, że dobrze rozumiesz samo pytanie i sposoby pomiarów. Dopiero po przełożeniu pytań na konkretne atrybuty (kolumny) danych możesz przeprowadzić skuteczną i sensowną analizę.

Kup książkę Poleć książkę

Skorowidz

Kup książkę Poleć książkę

E

F

Kup książkę Poleć książkę

Kup książkę Poleć książkę

G

H

I

J

Kup książkę Poleć książkę

K

L

Ł

M

Kup książkę Poleć książkę

N

O

P

Kup książkę Poleć książkę

R

Kup książkę Poleć książkę

S

Ś

T

Kup książkę Poleć książkę

U

V

W

Kup książkę Poleć książkę

Z

Ż

Kup książkę Poleć książkę


PROGRAM
PARTNERSKI
GRUPY HELION
1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ
Zmień swoją stronę WWW w działający bankomat!
Dowiedz się więcej i dołącz już dzisiaj!
http://program-partnerski.helion.pl
GRUPA
Helion

Po prostu R i dane. Wyciśniesz każdą kroplę wiedzy!

Aby surowe dane przekuć w gotową do wykorzystania wiedzę, potrzebna jest umiejętność ich analizy, przekształcania i niekiedy również wizualizacji. Nagrodą za włożony w to wysiłek jest lepsze rozumienie różnych złożonych zagadnień z wielu dziedzin wiedzy. Co więcej, znajomość procesów programowego przetwarzania danych pozwala na szybkie wykrywanie i opisywanie wzorców danych, praktycznie niemożliwych do dostrzeżenia innymi technikami. Dla wielu badaczy jednak barierą na drodze do skorzystania z tych atrakcyjnych możliwości jest konieczność pisania kodu.

Oto podręcznik programowania w języku R dla analityków danych, szczególnie przydatny dla osób, które nie mają doświadczenia w tej dziedzinie. Dokładnie opisano tu potrzebne narzędzia i technologie. Zamieszczono wskazówki dotyczące instalacji i konfiguracji oprogramowania do pisania kodu, wykonywania go i zarządzania nim, a także śledzenia wersji projektów i wprowadzanych w nich zmian oraz korzystania z innych podstawowych mechanizmów. Poszczególne kroki tworzenia kodu w języku R wyjaśniono dokładnie i przystępnie. Dzięki tej książce można płynnie przejść do konkretnych zadań i budować potrzebne aplikacje. Zrozumienie prezentowanych w niej treści ułatwiają liczne przykłady i ćwiczenia, co pozwala szybko przystąpić do skutecznego analizowania własnych zbiorów danych.

W książce:

- przygotowanie środowiska pracy i rozpoczęcie programowania w R
- podstawy zarządzania projektami, kontrola wersji i generowanie dokumentacji
- ramki danych, pakiety *dplyr* i *tidyr*
- kod do ich wizualizacji i pakiet *ggplot2*
- tworzenie aplikacji i techniki współpracy w zespołach specjalistów

MICHAEL FREEMAN – jest wykładowcą akademickim specjalizującym się w nauce o danych i ich wizualizacji danych. Wcześniej prowadził ogólnoświatowe badania dotyczące zdrowia publicznego. Interesuje się wykorzystaniem nauki o danych w obszarze sprawiedliwości społecznej.

JOEL ROSS – jest wykładowcą akademickim. Specjalizuje się w nauczaniu programowania. Interesuje się badaniami z zakresu gier i grywalizacji oraz systemami „przetwarzania bez granic". Prowadził też badania nad systemami finansowania społecznościowego i wspomagania ekorozwoju.

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

Sprawdź nasze szkolenia!
SZKOLENIA
AKADEMIA IT & BUSINESS
WWW.SZKOLENIA.HELION.PL

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-283-5782-2

9 788328 357822

INFORMATYKA W NAJLEPSZYM WYDANIU

Cena: 69,00 zł