MATT GOLDWASSER
UPOM MALIK
BENJAMIN JOHNSTON

SQL

ANALIZA DANYCH ZA POMOCĄ ZAPYTAŃ

WYDANIE II

WARSZTATY PRAKTYCZNE

Packt›

Helion

NAUKA O DANYCH I SZTUCZNA INTELIGENCJA

Tytuł oryginału: The Applied SQL Data Analytics Workshop: Develop your practical skills and prepare to become a professional data analyst, 2nd Edition

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-283-8474-3

Copyright © Packt Publishing 2020. First published in the English language under the title 'The Applied SQL Data Analytics Workshop - Second Edition – (9781800203679)'.

Polish edition copyright © 2022 by Helion S.A.
All rights reserved.

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 the Publisher.

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 wydawca 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 wydawca nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Helion S.A.
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/sqlan2
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

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

ROZDZIAŁ 1

Wprowadzenie do SQL-a dla analityków

Gdy zakończysz lekturę tego rozdziału, będziesz umiał opisywać dane i ich typy oraz kategoryzować dane na podstawie ich cech. Obliczysz podstawowe statystyki jednoczynnikowe dotyczące danych i zidentyfikujesz obserwacje odstające. Wykorzystasz też analizy dwuczynnikowe, które pomagają zrozumieć relacje między dwiema zmiennymi. Dowiesz się, do czego służy SQL, i zobaczysz, jak korzystać z tego języka w analityce danych. Na koniec nauczysz się podstaw relacyjnych baz danych i wykonywania operacji **CRUD** (od ang. *create*, *read*, *update* i *delete*, czyli tworzenie, wczytywanie, aktualizowanie i usuwanie) na tabelach.

Wprowadzenie

Dane zmieniły rzeczywistość w XXI wieku. Dzięki łatwemu dostępowi do komputerów firmy i inne organizacje mogły zmodyfikować sposób pracy z większymi i bardziej złożonymi zbiorami danych. Na podstawie danych można obecnie za pomocą kilku wierszy kodu komputerowego uzyskać informacje, jakie 50 lat temu były praktycznie niemożliwe do zdobycia. Dwa najważniejsze narzędzia w tej rewolucji to relacyjne bazy danych i ich podstawowy język, **SQL** (ang. *Structured Query Language*).

Choć teoretycznie możliwe jest analizowanie wszystkich danych ręcznie, komputery radzą sobie z tym zadaniem znacznie lepiej i są preferowanym narzędziem do przechowywania, porządkowania i przetwarzania danych. Do najważniejszych narzędzi związanych z danymi należą relacyjne bazy danych i język umożliwiający dostęp do nich, SQL. Te dwie technologie były przełomem w przetwarzaniu danych i nadal stanowią podstawowe narzędzia w większości firm korzystających z dużych ilości danych.

Kup książkę Poleć książkę

Firmy używają SQL-a jako podstawowej metody przechowywania większości danych. Ponadto dużą część tych danych umieszczają w specjalnych bazach nazywanych **hurtowniami danych** i **jeziorami danych**, które umożliwiają wykonywanie zaawansowanych analiz. Dostęp do prawie wszystkich hurtowni i jezior danych uzyskuje się za pomocą SQL-a. Dalej zobaczysz, jak używać SQL-a razem z analitycznymi platformami takimi jak hurtownie danych.

Zakładam, że wszyscy Czytelnicy mieli jakąś styczność z SQL-em. Jednak dla osób, które miały bardzo ograniczone doświadczenia z tym językiem lub dawno z niego nie korzystały, w tym rozdziale zamieszczam proste przypomnienie tego, czym są relacyjne bazy danych i SQL, a także podstawowy przegląd operacji i składni SQL-a. Omawiam też praktyczne ćwiczenia, które pomogą utrwalić te informacje.

Następny podrozdział pomoże Ci zrozumieć dane i ich typy.

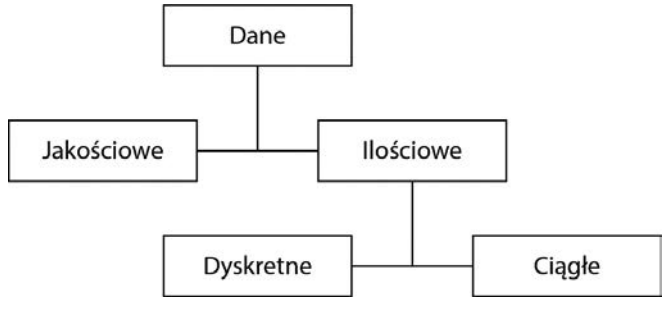
Świat danych

Zacznij od prostego pytania: czym są dane? **Dane** można traktować jako zapisane pomiary czegoś z rzeczywistego świata. Na przykład lista wzrostów to dane, ponieważ wzrost to pomiar odległości między głową a stopami danej osoby. Obiekt danych opisuje **jednostkę obserwacji**. W przypadku wzrostu jednostką obserwacji jest osoba.

Łatwo się domyślić, że istnieje dużo danych opisujących osobę: wiek, waga, czy pali papierosy itd. Jeden lub więcej pomiarów opisujących określoną jednostkę obserwacji to **punkt danych**, a każdy pomiar z punktu danych jest nazywany **zmienną** (lub **cechą**). Zestaw powiązanych punktów danych tworzy **zbiór danych**.

Rodzaje danych

Dane można podzielić na dwie podstawowe kategorie — **ilościowe** i **jakościowe** (rysunek 1.1).



Rysunek 1.1. Rodzaje danych

Kup książkę Poleć książkę

Dane ilościowe to pomiary, które można opisać za pomocą liczby. Dane jakościowe są reprezentowane za pomocą wartości nieliczbowych, na przykład tekstowo. Wzrost to dane ilościowe. Jednak opis kogoś jako „palącego" lub „niepalącego" to dane jakościowe.

Dane ilościowe można podzielić na dwie podkategorie: **dyskretne** i **ciągłe**. Dane ilościowe dyskretne to wartości mające określony poziom precyzji (zwykle mają postać liczb całkowitych). Na przykład liczba operacji w ciągu życia to dane dyskretne. Możesz mieć 0, 1 lub więcej operacji, ale nie możesz ich mieć 1,5. Zmienna ciągła teoretycznie może mieć dowolną precyzję. Na przykład masę ciała można opisać na dowolnym poziomie precyzji: 55 kg, 55,3 kg, 55,32 kg itd. W praktyce instrumenty pomiarowe oczywiście ograniczają precyzję, jeśli jednak wartość można przedstawić w bardziej precyzyjny sposób, zwykle zmienną uważa się za ciągłą.

Należy zauważyć, że dane jakościowe przeważnie można przekształcić na dane ilościowe, a dane ilościowe na dane jakościowe.

Rozważ to na przykładzie „palącego" i „niepalącego". Choć możesz stwierdzić, że należysz do jednej z tych kategorii, możesz też przedstawić je w inny sposób, jako ocenę stwierdzenia „palę regularnie", i użyć wartości logicznych `0` i `1` do reprezentowania odpowiedzi „prawda" i „fałsz".

Możliwe są też przekształcenia w odwrotną stronę, z danych ilościowych (na przykład wzrostu) na jakościowe. Zamiast myśleć o wzroście dorosłej osoby w kategoriach liczby cali lub centymetrów, możesz przypisać takie wartości do grup. Osoby wyższe niż 180 cm można uznać za „wysokie", osoby z przedziału od 160 do 180 cm za „średnie", a osoby niższe niż 160 cm za „niskie".

Analityka danych i statystyka

Surowe dane są jedynie zestawem wartości. W tej postaci nie są zbyt ciekawe. Dopiero gdy zaczniesz szukać wzorców w danych i je interpretować, możesz przejść do ciekawych czynności, takich jak prognozowanie przyszłości i identyfikowanie nieoczekiwanych zmian. Takie wzorce w danych to **informacje**. Duże uporządkowane kolekcje trwałych i rozbudowanych informacji oraz doświadczeń, które można wykorzystywać do opisywania i prognozowania zjawisk z rzeczywistego świata, zapewniają **wiedzę**. **Analiza danych** to proces przekształcania danych w informację, a dalej w wiedzę. Połączenie analizy danych z prognozowaniem to **analityka danych**.

Dostępnych jest wiele narzędzi, które pomagają zrozumieć dane. Jednym z najbardziej rozbudowanych narzędzi do analizy danych jest zastosowanie technik matematycznych do zbiorów danych. Jedną z tych technik matematycznych jest **statystyka**.

Kup książkę Poleć książkę

Rodzaje statystyki

Statystykę można podzielić na dwie kategorię: **statystykę opisową** i **wnioskowanie statystyczne**.

Statystyka opisowa służy do opisywania danych. Statystyki opisowe dotyczące jednej zmiennej ze zbioru danych to analizy **jednoczynnikowe**, a statystyki opisowe dotyczące jednocześnie dwóch lub więcej zmiennych to statystyki **wieloczynnikowe**.

We wnioskowaniu statystycznym zbiór danych jest traktowany jako **próbka**, czyli niewielka część pomiarów z większej grupy nazywanej **populacją**. Na przykład ankieta obejmująca 10 000 głosujących w wyborach krajowych daje próbkę opinii całej populacji głosującej w kraju. Wnioskowanie statystyczne służy do wyciągania wniosków na temat cech populacji na podstawie cech próbki.

> W tej książce skupiamy się przede wszystkim na statystykach opisowych. Więcej o wnioskowaniu statystycznym dowiesz się z podręczników do statystyki takich jak *Statistics* Davida Freedmana, Roberta Pisaniego i Rogera Purvesa.

Wyobraź sobie, że jesteś analitykiem zajmującym się polityką zdrowotną i otrzymałeś zbiór danych z informacjami o pacjentach (rysunek 1.2).

Rok Urodzenia	Kraj Urodzenia	Wzrost (cm)	Kolor Oczu	Liczba Wizyt Lekarza w 2018
1997	Egipt	182	Niebieskie	1
1988	Chiny	196	Piwne	2
1986	USA	180	Brązowe	2
1990	USA	166	Brązowe	1
1975	Indie	181	Zielone	3
1951	Niemcy	184	Brązowe	1
2000	Australia	174	Szare	5
1995	Indie	183	Brązowe	1
1992	Chiny	187	Brązowe	2
1987	USA	169	Niebieskie	2

Rysunek 1.2. Dane dotyczące opieki zdrowotnej

Gdy dostępny jest zbiór danych, często warto je poklasyfikować. Tu jednostką obserwacji dla zbioru danych jest pacjent, ponieważ każdy wiersz reprezentuje pojedynczą obserwację, która odpowiada unikatowemu pacjentowi. Trzy kolumny, `Data Urodzenia`, `Wzrost` i `Liczba Wizyt Lekarza`, są ilościowe, ponieważ do ich reprezentowania służą liczby. Dwie kolumny, `Kolor Oczu` i `Kraj Urodzenia`, są jakościowe.

Kup książkę Poleć książkę

Zadanie 1.01 — klasyfikowanie nowego zbioru danych

W tym zadaniu poklasyfikujesz dane ze zbioru. Niedługo zaczynasz pracę w startupie w innym mieście. Jesteś podekscytowany, ale przed przeprowadzką decydujesz się sprzedać wszystkie swoje rzeczy, w tym samochód. Nie jesteś pewien, jakiej ceny zażądać, dlatego chcesz zebrać trochę danych. Pytasz znajomych i członków rodziny, którzy ostatnio sprzedali samochód, o markę i cenę ich pojazdów. Na tej podstawie otrzymujesz zbiór danych z rysunku 1.3.

Data	Marka	Wartość Sprzedaży (w tysiącach zł)
01.02.2018	Ford	12
02.02.2018	Honda	15
02.02.2018	Mazda	19
03.02.2018	Ford	20
04.02.2018	Toyota	10
04.02.2018	Toyota	10
04.02.2018	Mercedes	30
05.02.2018	Ford	11
06.02.2018	Chevrolet	12,5
06.02.2018	Chevrolet	19

Rysunek 1.3. Dane na temat sprzedaży używanych samochodów

Oto kroki, jakie należy wykonać:

1. Ustalić jednostkę obserwacji.
2. Ocenić trzy kolumny pod kątem tego, czy zawierają dane ilościowe, czy jakościowe.
3. Przekształcić kolumnę `Marka` na kolumnę z danymi ilościowymi.

> Rozwiązanie tego zadania znajdziesz w „Dodatku”.

W tym zadaniu nauczyłeś się klasyfikować dane. W następnym podrozdziale poznasz różne metody z obszaru statystyki opisowej.

Metody z obszaru statystyki opisowej

Wcześniej wspomnieliśmy, że statystyka opisowa jest jednym ze sposobów na analizowanie danych w celu ich zrozumienia. Analiza jedno- i wieloczynnikowa mogą dać wgląd w dane zjawisko. W tym podrozdziale przyjrzysz się podstawowym technikom matematycznym, które możesz wykorzystać, aby lepiej zrozumieć i opisać zbiór danych.

Kup książkę Poleć książkę

Analiza jednoczynnikowa

Jedną z gałęzi statystyki jest analiza jednoczynnikowa. Jej metody pozwalają zrozumieć jedną zmienną ze zbioru danych. W tym punkcie omawiamy wybrane z najczęściej stosowanych technik analizy jednoczynnikowej.

Rozkład danych

Rozkład danych bazuje na liczbie określonych wartości w zbiorze danych. Przyjmijmy, że zbiór danych zawiera 1000 kart zdrowia, a jedną ze zmiennych w tych kartach jest kolor oczu. Jeśli po przejrzeniu tego zbioru stwierdzisz, że 700 osób ma brązowe oczy, 200 osób ma zielone oczy, a 100 — niebieskie, opiszesz rozkład danych, a dokładniej **rozkład bezwzględnej częstości występowania**. Gdybyśmy podali nie liczbę wystąpień wartości w zbiorze danych, lecz odsetek jej wystąpień w łącznej liczbie punktów danych, opisalibyśmy **rozkład względnej częstości występowania**. W przykładzie z kolorami oczu rozkład względnej częstości występowania to: oczy brązowe 70%, oczy zielone 20%, oczy niebieskie 10%.

Łatwo jest obliczyć rozkład, gdy zmienna przyjmuje niewielką liczbę stałych wartości takich jak kolor oczu. Co jednak ze zmiennymi ilościowymi, które mogą przyjmować wiele różnych wartości, na przykład ze wzrostem? Ogólna metoda obliczania rozkładu dla zmiennych tego rodzaju polega na tworzeniu „kubełków", do których można przypisać poszczególne wartości, i obliczaniu rozkładów na podstawie tych kubełków. Na przykład wzrost można podzielić na kubełki obejmujące po 5 cm, aby uzyskać rozkład bezwzględnej częstości występowania. Następnie można podzielić każdy wiersz tabeli przez łączną liczbę punktów danych (10 000) i otrzymać rozkład względnej częstości występowania.

Inną przydatną techniką jest tworzenie wykresów rozkładów. Teraz utworzysz **histogram**, który jest graficzną reprezentacją rozkładu ciągłego z użyciem kubełków.

Ćwiczenie 1.01 — tworzenie histogramu

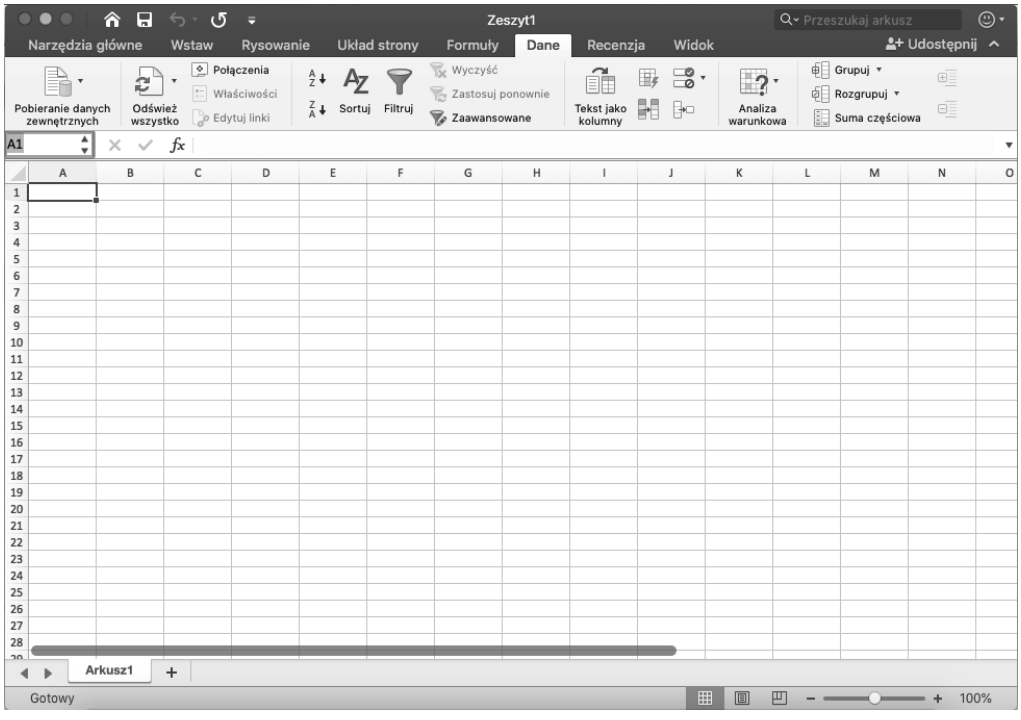
W tym ćwiczeniu użyjesz programu Microsoft Excel do utworzenia histogramu. Przyjmij, że jesteś analitykiem polityki zdrowotnej i chcesz zobaczyć rozkład wzrostu, aby dostrzec w nim wzorce. Aby wykonać to zadanie, musisz przygotować histogram.

> Do tworzenia histogramów możesz użyć arkusza kalkulacyjnego, na przykład w programie Excel, albo języków takich jak Python lub R. Dla wygody tu posłużysz się Excelem. Zbiory danych z tego rozdziału znajdziesz w serwisie GitHub: *https://packt.live/2B1apb3*.

Wykonaj następujące kroki:

1. Otwórz pusty skoroszyt w programie Microsoft Excel (rysunek 1.4).
2. Przejdź do zakładki *Dane* i wybierz opcję *Pobieranie danych zewnętrznych/ Z pliku tekstowego*.

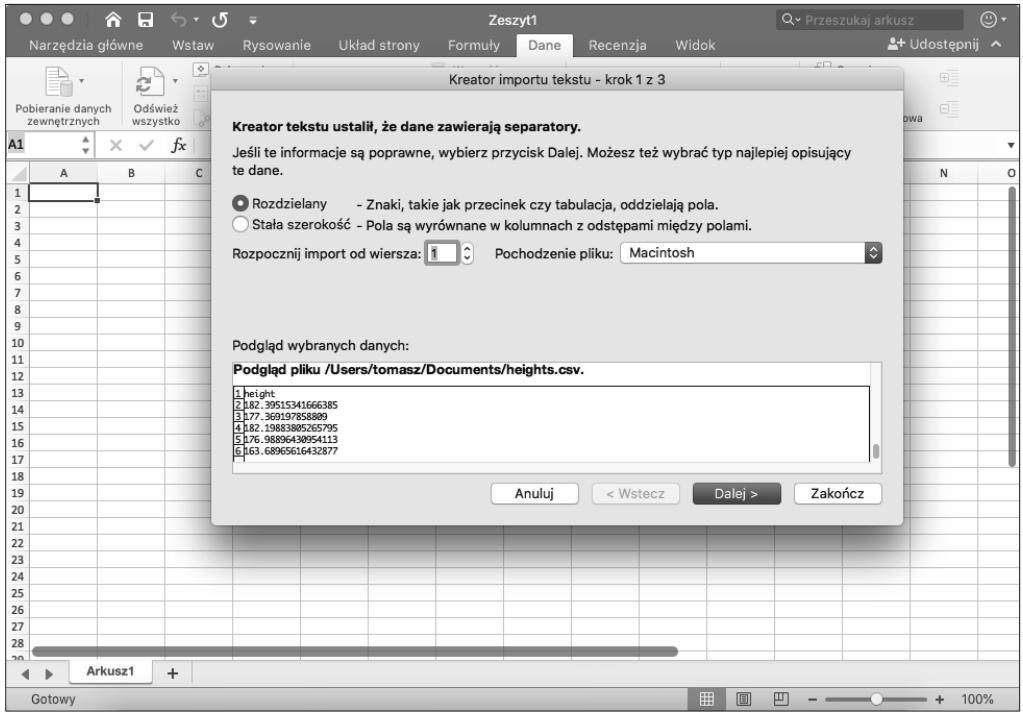
Kup książkę Poleć książkę



Rysunek 1.4. Pusty skoroszyt w Excelu

3. Znajdź plik ze zbiorem danych *heights.csv* w katalogu *Datasets* z repozytorium serwisu GitHub. Po wskazaniu tego pliku kliknij *OK*.
4. Wybierz opcję *Rozdzielany* w oknie dialogowym *Kreator importu tekstu*. Import należy rozpocząć od wiersza *1*. Następnie kliknij *Dalej*.
5. Wybierz ogranicznik używany w pliku. Ponieważ ten plik ma tylko jedną kolumnę, nie występują w nim ograniczniki. W plikach CSV tradycyjnie ogranicznikami są przecinki; w przyszłości używaj ograniczników odpowiednich dla zbiorów danych. Teraz kliknij *Dalej*.
6. Wybierz *Ogólne* w sekcji *Format danych kolumny*. Kliknij przycisk *Zaawansowane* i w nowym oknie jako *Separator dziesiętny* wybierz kropkę, po czym kliknij *OK*. Następnie kliknij *Zakończ*.
7. W oknie dialogowym z pytaniem *Gdzie chcesz umieścić dane?* wybierz opcję *Istniejący arkusz* i nie zmieniaj wartości w polu tekstowym obok tej opcji. Kliknij przycisk *OK*.
8. W kolumnie `C` zapisz liczby `140`, `145`, `150` itd., zwiększając wartości o 5 aż do `220`. Umieść je w komórkach od `C2` do `C18`, tak jak na rysunku 1.6.

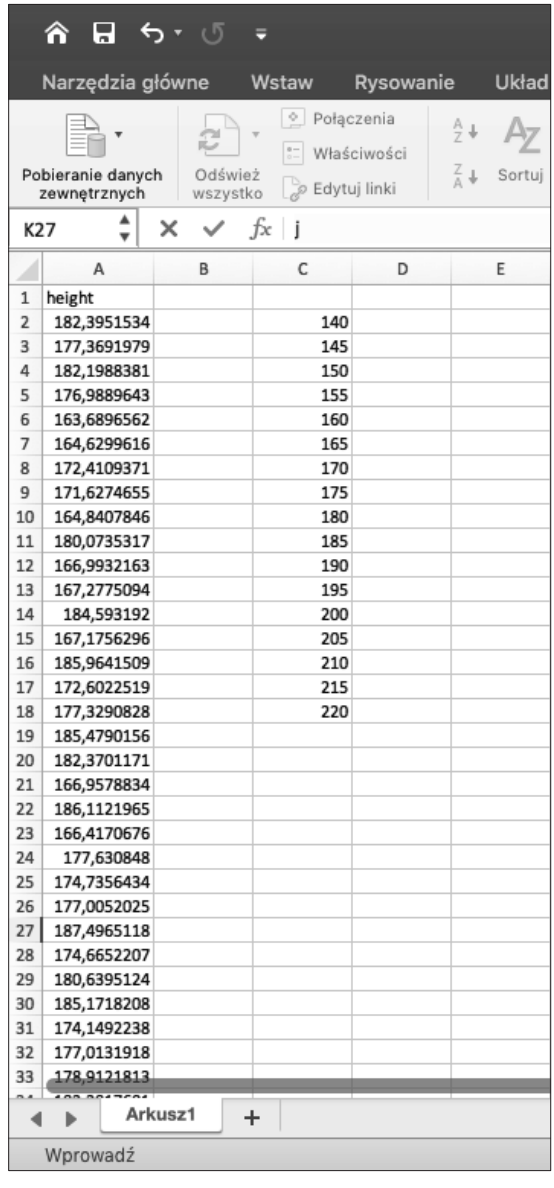
Kup książkę Poleć książkę



Rysunek 1.5. Zaznacz opcję Rozdzielany

9. W zakładce *Dane* wybierz opcję *Analiza danych* (jeśli jej nie widzisz, zastosuj się do instrukcji ze strony *https://support.office.com/en-us/article/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4*, aby ją dodać).
10. W polu wyboru, które się pojawi, wybierz opcję *Histogram* i kliknij przycisk *OK*.
11. Aby podać opcję *Zakres komórek*, kliknij przycisk po prawej stronie pola tekstowego. Powinieneś wrócić do arkusza *Arkusz1*, gdzie widoczne będzie puste pole z przyciskiem z czerwoną strzałką. Przeciągnij kursor i zaznacz wszystkie dane w arkuszu od komórki `A2` do `A10001`. Następnie kliknij przycisk z czerwoną strzałką.
12. Aby ustawić opcję *Zakres zbioru*, kliknij przycisk po prawej stronie pola tekstowego. Powinieneś wrócić do arkusza *Arkusz1*, gdzie widoczne będzie puste pole z przyciskiem z czerwoną strzałką. Przeciągnij kursor i zaznacz wszystkie dane w arkuszu od komórki `C2` do `C18`. Następnie kliknij przycisk z czerwoną strzałką.
13. W sekcji *Opcje wyjścia* zaznacz pole *Nowy arkusz* i upewnij się, że zaznaczone jest pole *Wykres wyjściowy*, tak jak na rysunku 1.7. Następnie kliknij przycisk *OK*.

Kup książkę Poleć książkę



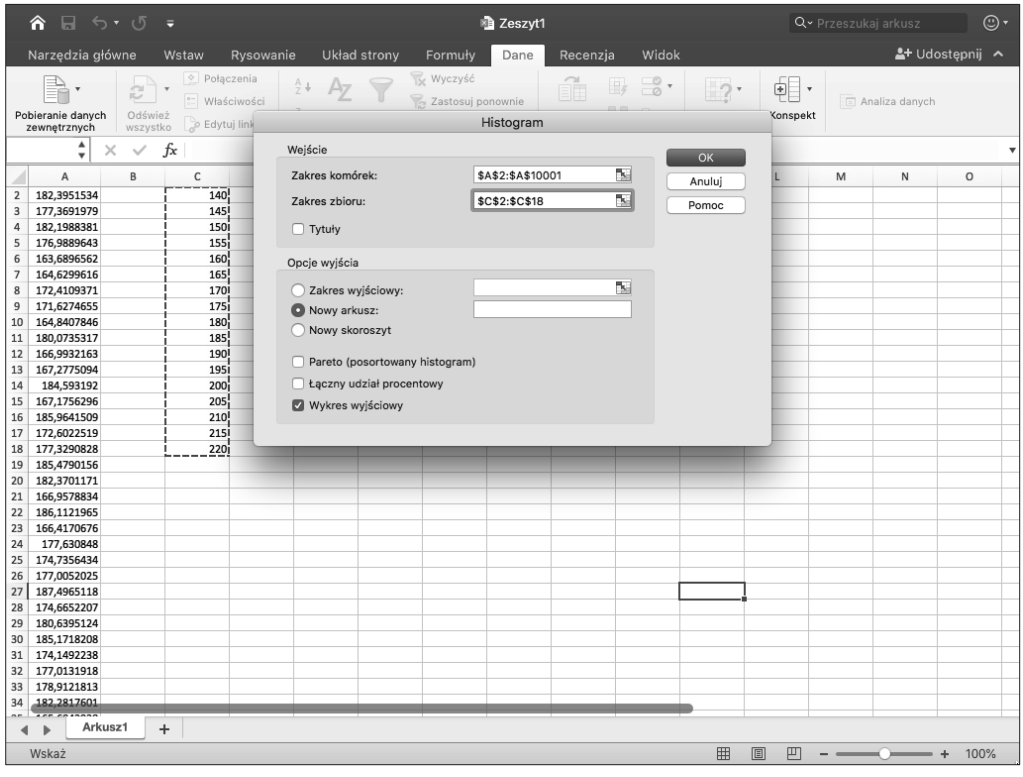
	A	B	C	D	E
1	height				
2	182,3951534		140		
3	177,3691979		145		
4	182,1988381		150		
5	176,9889643		155		
6	163,6896562		160		
7	164,6299616		165		
8	172,4109371		170		
9	171,6274655		175		
10	164,8407846		180		
11	180,0735317		185		
12	166,9932163		190		
13	167,2775094		195		
14	184,593192		200		
15	167,1756296		205		
16	185,9641509		210		
17	172,6022519		215		
18	177,3290828		220		
19	185,4790156				
20	182,3701171				
21	166,9578834				
22	186,1121965				
23	166,4170676				
24	177,630848				
25	174,7356434				
26	177,0052025				
27	187,4965118				
28	174,6652207				
29	180,6395124				
30	185,1718208				
31	174,1492238				
32	177,0131918				
33	178,9121813				

Rysunek 1.6. Wprowadzanie danych w arkuszu Excela

14. Kliknij *Arkusz2*. Znajdź wykres i kliknij dwukrotnie tytuł *Histogram*. Zastąp go słowem *Wzrost*. Powinieneś uzyskać wykres podobny do tego z rysunku 1.8.

Przyjrzenie się kształtowi rozkładu pozwala odkryć ciekawe wzorce. Zauważ, że ten rozkład ma symetryczny kształt dzwona. Jest to tak zwany *rozkład normalny*, występujący w wielu zbiorach danych. W tej książce nie omawiamy szczegółowo tego rozkładu, ale zwróć na niego uwagę w trakcie analiz danych; często będziesz na niego natrafiać.

Kup książkę Poleć książkę



Rysunek 1.7. Wybierz opcję Nowy arkusz

Rysunek 1.8. Rozkład wzrostu dorosłych mężczyzn

Kwantyle

Jednym ze sposobów na liczbowy opis rozkładu jest użycie kwantyli. Kwantyle rzędu N to zbiór $n - 1$ punktów dzielących zmienną na n grup. Te punkty czasem nazywa się **punktami podziału**. Na przykład kwantyle rzędu 4 (nazywane kwartylami) to grupa trzech punktów, które dzielą zmienną na cztery w przybliżeniu równe grupy wartości. Na rysunku 1.9 wymienione są nazwy kwantyli różnego rzędu.

Kup książkę Poleć książkę

N	Nazwa
3	Tercyle
4	Kwartyle
5	Kwintyle
10	Decyle
20	Vingtyle
100	Percentyle

Rysunek 1.9. Nazwy kwantyli rzędu N

Istnieją różne procedury obliczania kwantyli. Tu użyjesz następującej techniki do obliczania kwantyli rzędu N dla *d* punktów danych dla jednej zmiennej:

1. Uporządkuj punkty danych od najmniejszego do największego.
2. Ustal liczbę *n* dla kwantyli rzędu N, jakie chcesz wyznaczyć, oraz liczbę punktów podziału $(n - 1)$.
3. Ustal liczbę *k*-tego punktu podziału, jaki chcesz obliczyć, czyli liczbę z przedziału od 1 do $n - 1$. Jeśli rozpoczynasz obliczenia, użyj *k* równego 1.
4. Wyznacz indeks *i* dla *k*-tego punktu podziału. Użyj wzoru z rysunku 1.10.

$$i = \left\lfloor \frac{k}{n}(d-1) \right\rfloor + 1$$

Rysunek 1.10. Indeks

5. Jeśli *i* jest liczbą całkowitą, wybierz element o tym numerze spośród uporządkowanych punktów danych. Jeżeli *k*-ty punkt podziału nie jest liczbą całkowitą, znajdź elementy bezpośrednio przed oraz po *i*. Oblicz różnicę między wartościami tych elementów i pomnóż ją przez część dziesiętną uzyskanego indeksu. Dodaj wynik do mniejszego z dwóch uwzględnianych elementów.
6. Powtarzaj kroki od 2. do 5. dla różnych wartości *k* do momentu obliczenia wszystkich punktów podziału.

Te kroki dość trudno jest zrozumieć bez kontekstu, dlatego warto wykonać ćwiczenie. Dzięki wielu współczesnym narzędziom, między innymi SQL-owi, komputery mogą szybko obliczać kwantyle za pomocą wbudowanych mechanizmów.

Ćwiczenie 1.02 — obliczanie kwartyli dla sprzedaży dodatków

W tym ćwiczeniu za pomocą Excela poklasyfikujesz dane i obliczysz kwartyle dotyczące sprzedaży samochodów. Twój nowy szef chce, abyś przejrzał dane, zanim zaczniesz pracę w poniedziałek, co pozwoli Ci lepiej zrozumieć jedno z zadań, jakimi będziesz się zajmować

Kup książkę Poleć książkę

— zwiększenie sprzedaży dodatkowego wyposażenia przy zakupie samochodów. Szef przesyła Ci listę 11 transakcji i kwot wydanych na dodatki do podstawowej wersji nowego modelu ZoomZoom Chi. Oto wartości sprzedaży dodatkowego wyposażenia: 5000, 1700, 8200, 1500, 3300, 9000, 2000, 0, 0, 2300, 4700.

> Wszystkie zbiory danych używane w tym rozdziale znajdziesz w serwisie GitHub: *https://packt.live/2B1apb3*.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Otwórz pusty skoroszyt w programie Microsoft Excel.
2. Przejdź do zakładki *Dane* i wybierz opcję *Pobieranie danych zewnętrznych/ Z pliku tekstowego*.
3. Plik ze zbiorem danych *auto_upgrades.csv* znajdziesz w katalogu *Datasets* w repozytorium w serwisie GitHub. Przejdź do tego pliku i kliknij przycisk *Pobierz dane*.
4. Wybierz opcję *Rozdzielany* w oknie dialogowym *Kreator importu tekstu*. Import musi się rozpoczynać od wiersza *1*. Następnie kliknij *Dalej*.
5. Wybierz ogranicznik używany w pliku. Ponieważ ten plik ma tylko jedną kolumnę, nie występują w nim ograniczniki. W plikach CSV tradycyjnie ogranicznikami są przecinki; w przyszłości używaj ograniczników odpowiednich dla zbiorów danych. Teraz kliknij *Dalej*.
6. Wybierz *Ogólne* w sekcji *Format danych kolumny*. Kliknij przycisk *Zaawansowane* i w nowym oknie jako *Separator dziesiętny* wybierz kropkę, po czym kliknij *OK*. Następnie kliknij *Zakończ*.
7. W oknie dialogowym z pytaniem *Gdzie chcesz umieścić dane?* wybierz opcję *Istniejący arkusz* i nie zmieniaj wartości w polu tekstowym obok tej opcji. Kliknij przycisk *OK*.
8. Kliknij komórkę *A1*, a następnie otwórz zakładkę *Dane* i wybierz opcję *Sortuj*.
9. Pojawi się okno dialogowe sortowania. Kliknij przycisk *OK*. Wartości zostaną posortowane od najmniejszych do największych. Lista z rysunku 1.11 przedstawia posortowane wartości.
10. Teraz ustal liczbę *kwantyli rzędu n* i punktów podziału, jakie musisz obliczyć. Kwartyle to kwantyle rzędu 4, zgodnie z rysunkiem 1.9. Ponieważ liczba punktów podziału jest o 1 mniejsza od liczby *kwantyli rzędu n*, wiadomo, że potrzebne są trzy punkty podziału.
11. Oblicz indeks pierwszego punktu podziału. Tu $k=1$, liczba wartości w populacji (d) jest równa *11*, a liczba *kwantyli rzędu n* (n) to cztery. Po podstawieniu tych wartości w równaniu z rysunku 1.12 uzyskasz *3,5*.

Kup książkę Poleć książkę

	A	B	C	D
1	Add-on Sales ($)			
2	0			
3	0			
4	1500			
5	1700			
6	2000			
7	2300			
8	3300			
9	4700			
10	5000			
11	8200			
12	9000			
13				

Rysunek 1.11. Posortowane wartości sprzedaży dodatkowego wyposażenia

$$i = \left[\frac{k}{n}(d-1)\right] + 1$$
$$i = \left[\frac{1}{4}(11-1)\right] + 1$$
$$i = \frac{10}{4} + 1$$
$$i = \frac{10}{4} + 1$$
$$i = 2{,}5 + 1 = 3{,}5$$

Rysunek 1.12. Obliczanie indeksu pierwszego punktu podziału

12. Ponieważ indeks 3,5 nie jest liczbą całkowitą, najpierw trzeba znaleźć elementy trzeci i czwarty (1500 i 1700). Należy ustalić różnicę między nimi (200), a następnie pomnożyć ją przez część dziesiętną indeksu, czyli 0,5, co daje w wyniku 100. Tę wartość należy dodać do trzeciego elementu (1500), uzyskasz więc wartość 1600.
13. Powtórz kroki od 2. do 5. procedury dla $k=2$ i $k=3$, aby obliczyć poziom drugiego i trzeciego kwartyla. Powinieneś otrzymać wartości 2300 i 4850.

W tym ćwiczeniu zobaczyłeś, jak klasyfikować dane i obliczać kwartyle za pomocą Excela.

Kup książkę Poleć książkę

Tendencja centralna

Jedno z podstawowych pytań na temat zmiennych ze zbioru danych dotyczy typowej wartości określonej zmiennej. Ta wartość jest często nazywana **tendencją centralną** zmiennej. Do opisywania tendencji centralnej można użyć wielu wartości obliczanych na podstawie zbioru danych; wszystkie one mają wady i zalety. Oto niektóre miary tendencji centralnej:

- **Wartość modalna** — jest to wartość, która najczęściej występuje w rozkładzie zmiennej. Na rysunku 1.2 dla przykładu dotyczącego koloru oczu wartość modalna to „oczy brązowe", ponieważ występuje ona w tym zbiorze najczęściej. Jeśli istnieje kilka takich wartości, zmienna jest nazywana **wielomodalną** i należy podać wszystkie najczęściej występujące wartości. Jeżeli żadna wartość się nie powtarza, w danym zbiorze wartość modalna nie występuje. Wartość modalna jest przydatna, jeśli zmienna może przyjmować niewielką, stałą liczbę wartości. Trudno jest ją jednak wyznaczyć dla ciągłych zmiennych ilościowych, na przykład dla wzrostu. Wtedy tendencję centralną lepiej jest określać za pomocą innych miar.
- **Średnia** — średnia dla zmiennej to wartość obliczona przez zsumowanie wszystkich jej wartości i podzielenie wyniku przez liczbę punktów danych. Załóżmy, że używamy niewielkiego zbioru danych z wiekiem: 26, 25, 31, 35 i 29. Średnia dla tego zbioru wynosi 29,2, ponieważ ten właśnie wynik uzyskasz, gdy zsumujesz tych pięć liczb, a następnie podzielisz uzyskaną sumę przez 5 (czyli liczbę punktów danych). Średnią łatwo jest obliczyć i zwykle dobrze opisuje ona „typową" wartość zmiennej. Nie jest więc zaskoczeniem, że stanowi ona jedną z najczęściej podawanych statystyk opisowych w literaturze przedmiotu. Jednak średnia jako miara tendencji centralnej ma ważną wadę: jest wrażliwa na **wartości odstające**.

  Wartość odstająca to punkt danych, który znacznie różni się od pozostałych danych i występuje bardzo rzadko. Wartości odstające często można identyfikować za pomocą technik graficznych, na przykład na wykresach punktowych lub skrzynkowych, gdzie widoczne są wszystkie punkty danych bardzo oddalone od pozostałych. Gdy w zbiorze danych występuje wartość odstająca, można go nazwać **skośnym zbiorem danych**. Częste powody występowania wartości odstających to nieoczyszczone dane, niezwykle rzadkie zdarzenia i problemy z instrumentami pomiarowymi. Wartości odstające często zniekształcają średnią do tego stopnia, że przestaje ona reprezentować typowe wartości danych.
- **Mediana** — jest nazywana także drugim kwartylem lub percentylem 50% i stanowi dość dziwną miarę tendencji centralnej, jednak ma ważne zalety w porównaniu ze średnią. Aby obliczyć medianę, posortuj wartości zmiennej od najmniejszej do największej, a następnie wybierz wartość środkową. Dla nieparzystej liczby punktów danych jest to środkowa wartość w uporządkowanych danych. Gdy liczba punktów danych jest parzysta, użyj średniej dwóch środkowych wartości.

  Choć obliczanie mediany jest trochę niewygodne, jest ona mniej wrażliwa na wartości odstające (w porównaniu ze średnią). Aby się o tym przekonać, oblicz medianę dla skośnego zbioru danych z wiekiem: 26, 25, 31, 35, 29 i 82. Mediana dla tego zbioru danych wynosi 30. Wartość ta jest dużo bliższa typowej wartości

Kup książkę Poleć książkę

tego zbioru danych niż średnia, która wynosi 38. Ta odporność na wartości odstające jest jednym z głównych powodów używania mediany.

Zgodnie z ogólną regułą warto obliczyć zarówno średnią, jak i medianę zmiennej. Jeśli te miary znacznie się od siebie różnią, w zbiorze danych mogą występować wartości odstające.

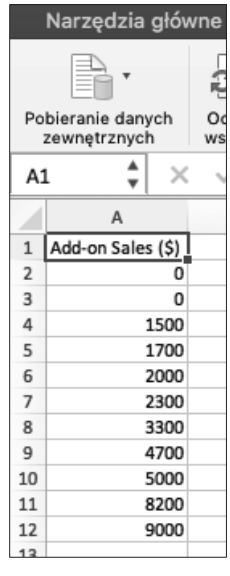
W następnym ćwiczeniu zobaczysz, jak obliczać miary tendencji centralnej.

Ćwiczenie 1.03 — obliczanie miar tendencji centralnej dla sprzedaży dodatków

W tym ćwiczeniu obliczysz miary tendencji centralnej dla danych w Excelu. Aby lepiej zrozumieć dane dotyczące sprzedaży dodatkowego wyposażenia (które można dokupić do podstawowego modelu), należy ustalić typową wartość zmiennej. Oblicz wartość modalną, średnią i medianę dla tych danych. Oto wartość dodatków w 11 transakcjach zakupu samochodów: 5000, 1700, 8200, 1500, 3300, 9000, 2000, 0, 0, 2300 i 4700.

W tym ćwiczeniu wykonaj następujące kroki:

1. Najpierw oblicz wartość modalną, aby wyznaczyć najczęściej występującą wartość. Ponieważ w tym zbiorze danych najczęściej pojawia się 0, wartość modalna to *0*.
2. Teraz oblicz średnią. Zsumuj liczby z kolumny `Add-on Sales`. Wynik to 37 700. Podziel go przez liczbę wartości (11), a otrzymasz średnią 3427,27.
3. W ostatnim kroku oblicz medianę. W tym celu posortuj dane tak jak na rysunku 1.13.



Rysunek 1.13. Posortowane dane o sprzedaży dodatków

Kup książkę Poleć książkę

Ustal środkową wartość. Ponieważ jest 11 punktów danych, środkowa jest szósta wartość na liście. Sprawdź więc szósty element w posortowanych danych, a otrzymasz medianę równą 2300.

Po zapoznaniu się z tendencją centralną możesz przejść do innej cechy danych — dyspersji.

Gdy porównasz średnią z medianą, zobaczysz, że znacznie się od siebie różnią. Wcześniej wspomnieliśmy, że jest to oznaka występowania wartości odstających w zbiorze danych. Dalej dowiesz się, jak stwierdzić, które wartości są odstające.

Dyspersja

Inną ciekawą cechą zbioru danych jest to, jak blisko siebie znajdują się wartości zmiennej w punktach danych. Na przykład zbiory liczb [100, 100, 100] i [50, 100, 150] oba mają średnią 100, ale wartości w drugim są bardziej rozproszone niż w pierwszym. Cecha opisująca rozproszenie danych jest nazywana **dyspersją**.

Istnieje wiele sposobów pomiaru dyspersji zmiennej. Oto kilka popularnych technik:

- **Rozstęp** — jest to różnica między największą a najmniejszą wartością zmiennej. Można go bardzo łatwo obliczyć, ale jest niezwykle wrażliwy na wartości odstające. Ponadto nie daje informacji na temat rozproszenia wartości pośrodku zbioru danych.
- **Odchylenie standardowe i wariancja** — odchylenie standardowe to pierwiastek kwadratowy ze średniej kwadratów różnic wartości punktu danych od średniej. Odchylenie standardowe przyjmuje wartości od zera do dodatniej nieskończoności. Im bliżej odchylenie standardowe jest zera, tym mniejsze są różnice między liczbami w zbiorze danych. Gdy odchylenie standardowe wynosi zero, wszystkie wartości w zbiorze danych są takie same.

  Warto zwrócić uwagę na to, że są dwa różne wzory na obliczanie odchylenia standardowego, co ilustruje rysunek 1.14. Gdy zbiór danych reprezentuje całą populację, należy obliczyć odchylenie standardowe dla populacji, używając wzoru A z tego rysunku. Jeśli próbka reprezentuje tylko część obserwacji, należy posłużyć się wzorem B do obliczenia odchylenia standardowego dla próbki. Gdy masz wątpliwości, zastosuj wzór na odchylenie standardowe dla próbki, ponieważ jest on bardziej zachowawczy. W praktyce gdy liczba punktów danych jest duża, różnica między tymi dwoma wzorami jest bardzo niewielka.

A) $\sqrt{\frac{\sum_{i=1}^{n}(x_i - u_x)^2}{n}}$ B) $\sqrt{\frac{\sum_{i=1}^{n}(x_i - u_x)^2}{n-1}}$

Rysunek 1.14. Wzory na odchylenie standardowe dla populacji (A) i próbki (B)

Kup książkę Poleć książkę

Najczęściej używaną miarą do opisywania dyspersji jest właśnie odchylenie standardowe. Jednak, podobnie jak rozstęp, jest ona wrażliwa na wartości odstające, choć w mniejszym stopniu. Obliczanie odchylenia standardowego jest też dość skomplikowane, lecz współczesne narzędzia zwykle umożliwiają łatwe uzyskanie tej miary.

Na koniec ostatnia uwaga: czasem możesz zetknąć się z powiązaną wartością, wariancją. Równa się ona odchyleniu standardowemu podniesionemu do kwadratu.

- **Rozstęp ćwiartkowy** — jest to różnica między pierwszym kwartylem (Q1, nazywany też dolnym) i trzecim kwartylem (Q3, nazywany też górnym).

> Więcej informacji o obliczaniu kwantyli i kwartyli znajdziesz w punkcie „Rozkład danych" w tym rozdziale.

Rozstęp ćwiartkowy, w odróżnieniu od rozstępu i odchylenia standardowego, jest odporny na wartości odstające. Dlatego choć jest to jedna z najtrudniejszych do obliczenia miar, dobrze nadaje się do pomiaru dyspersji zbioru danych. Często jest też używana do definiowania wartości odstających. Jeśli wartość w zbiorze danych jest mniejsza niż (Q1 – 1,5 × rozstęp ćwiartkowy) lub większa niż (Q1 + 1,5 × rozstęp ćwiartkowy), jest uznawana za odstającą.

Aby lepiej zrozumieć dyspersję, wykonaj następne ćwiczenie.

Ćwiczenie 1.04 — obliczanie dyspersji dla sprzedaży dodatków

W tym ćwiczeniu obliczysz rozstęp, odchylenie standardowe i rozstęp ćwiartkowy. Aby lepiej zrozumieć sprzedaż dodatków i opcjonalnych urządzeń, dokładnie przyjrzyj się dyspersji danych. Oto dane dla 11 transakcji zakupu dodatkowego wyposażenia: 5000, 1700, 8200, 1500, 3300, 9000, 2000, 0, 0, 2300 i 4700.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Oblicz zakres. W tym celu znajdź najmniejszą wartość w danych (0) i odejmij ją od wartości maksymalnej (9000). Uzyskasz wynik 9000.
2. Obliczenie odchylenia standardowego wymaga, aby najpierw ustalić, czy ma ono dotyczyć próbki czy populacji. Ponieważ 11 analizowanych punktów danych reprezentuje niewielką próbkę wszystkich transakcji, obliczysz odchylenie standardowe dla próbki.
3. Następnie znajdź średnią zbioru danych. Obliczyłeś ją już w „Ćwiczeniu 1.02 — obliczanie kwartyli dla sprzedaży dodatków"; wynik to 3427,27.
4. Teraz odejmij wszystkie punkty danych od średniej i podnieś wynik do kwadratu. Wyniki są pokazane na rysunku 1.15.

Kup książkę Poleć książkę

Add-on Sales ($)	Różnica względem średniej	Kwadrat różnicy względem średniej
5000	1572,727273	2473471,074
1700	-1727,272727	2983471,074
8200	4772,727273	22778925,62
1500	-1927,272727	3714380,165
3300	-127,2727273	16198,34711
9000	5572,727273	31055289,26
2000	-1427,272727	2037107,438
0	-3427,272727	11746198,35
0	-3427,272727	11746198,35
2300	-1127,272727	1270743,802
4700	1272,727273	1619834,711

Rysunek 1.15. Obliczanie sumy kwadratów różnic

5. Zsumuj wartości z kolumny *Kwadrat różnicy względem średniej*; wynik to 91 441 818.
6. Podziel tę sumę przez liczbę punktów danych minus 1 (czyli przez 10) i wyciągnij pierwiastek kwadratowy. Te obliczenia powinny dać odchylenie standardowe dla próbki równe 3023,93.
7. Aby obliczyć rozstęp ćwiartkowy, wyznacz pierwszy i trzeci kwartyl. Obliczenia znajdziesz w „Ćwiczeniu 1.02 — obliczanie kwartyli dla sprzedaży dodatków"; wyniki to 1600 i 4850. Odejmij te dwie wartości, a uzyskasz wynik 3250.

W tym ćwiczeniu obliczyłeś rozstęp, odchylenie standardowe i rozstęp ćwiartkowy. W następnym ćwiczeniu zobaczysz, jak posłużyć się analizą dwuczynnikową do znajdowania wzorców.

Analiza dwuczynnikowa

Do tej pory omawialiśmy metody opisu jednej zmiennej. Teraz zobaczysz, jak za pomocą analizy dwuczynnikowej wykrywać wzorce dotyczące dwóch zmiennych.

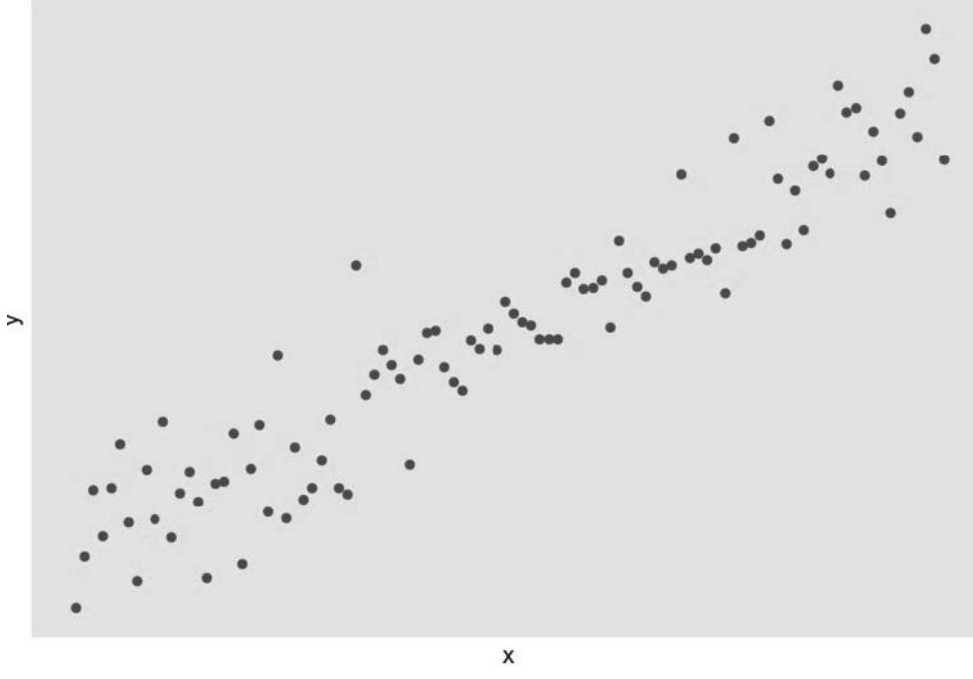
Wykresy punktowe

Ogólną regułą, jaką odkryjesz w analityce, jest to, że wykresy są niezwykle pomocne w wyszukiwaniu wzorców. Podobnie jak histogramy pomagają zrozumieć jedną zmienną, wykresy punktowe ułatwiają zapoznanie się z dwiema zmiennymi. Wykresy punktowe można łatwo przygotować za pomocą wybranego arkusza kalkulacyjnego.

> Wykresy punktowe są pomocne przede wszystkim w sytuacji, gdy liczba punktów jest niewielka (zwykle od 30 do 500). Jeśli liczba punktów jest duża i naniesienie ich na wykres kończy się uzyskaniem jednej wielkiej plamy, wybierz losową próbkę 200 punktów i utwórz wykres na ich podstawie, co może pomóc Ci w wykryciu ciekawych trendów.

Kup książkę Poleć książkę

Na wykresie punktowym możesz znaleźć wiele różnych wzorców. Najczęściej wyszukiwane są trendy wzrostowe i malejące dla dwóch zmiennych. Pozwala to stwierdzić, czy wraz ze wzrostem wartości jednej zmiennej druga zmienna też rośnie, czy maleje. Trend wskazuje na to, że między dwiema zmiennymi może występować przewidywalna zależność matematyczna. Istnieje na przykład trend rosnący dla zależności między wiekiem i zarobkami. Rysunek 1.16 ilustruje przykładowy trend liniowy.



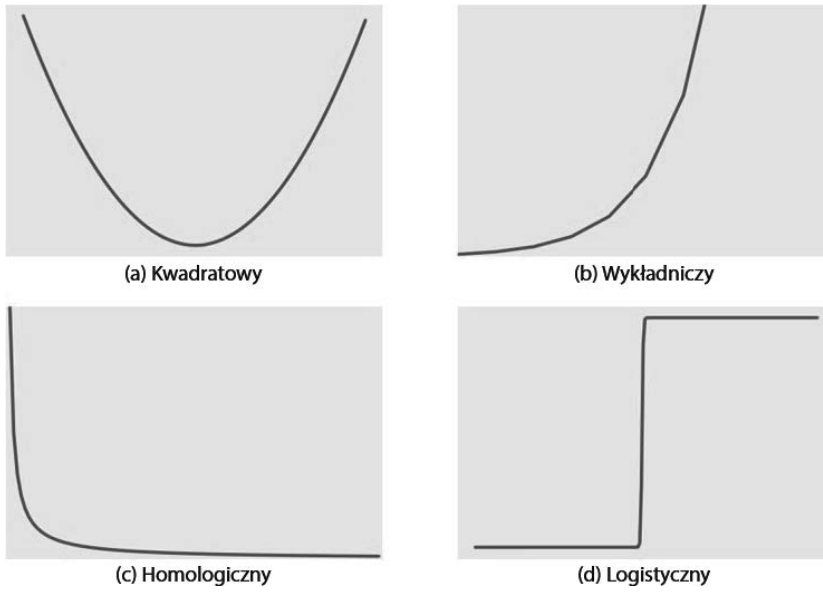
Rysunek 1.16. Rosnący trend liniowy dla dwóch zmiennych — wieku i zarobków osób

Istnieje też wiele wartych uwagi trendów nieliniowych, w tym kwadratowe, wykładnicze, homologiczne i logistyczne. Na rysunku 1.17 pokazane są niektóre z nich.

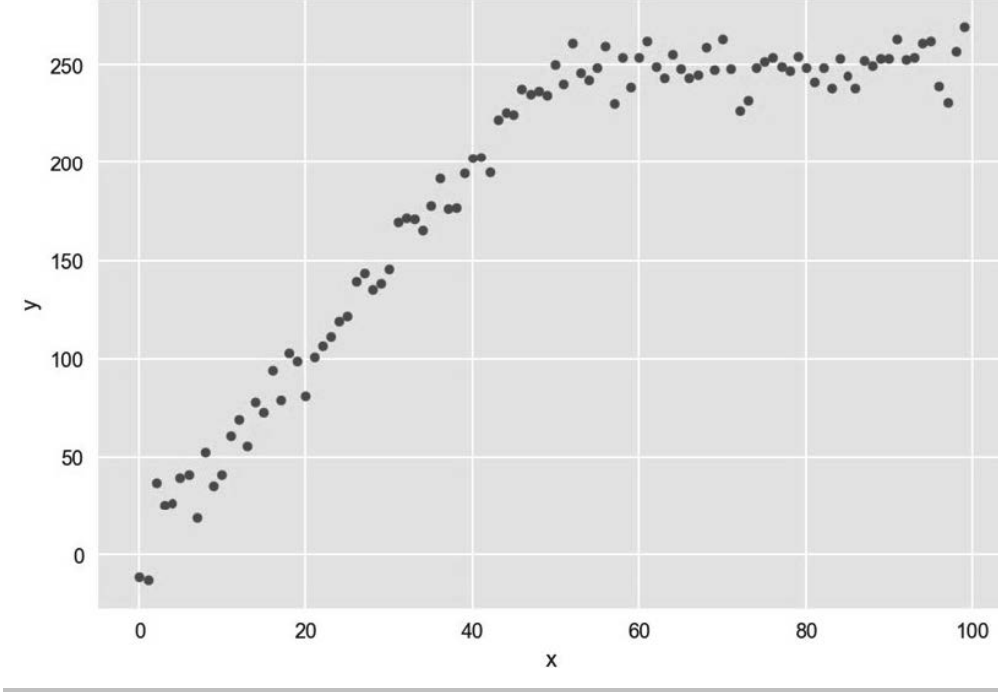
> Proces przybliżonego opisu trendu za pomocą funkcji matematycznej jest nazywany **analizą regresji**. Jest ona bardzo ważna w analityce, ale jej omawianie wykracza poza zakres tej książki. Więcej informacji o analizie regresji znajdziesz w zaawansowanych książkach, takich jak *Regression Modeling Strategies* Franka E. Harrela Jr.

Choć trendy pomagają w zrozumieniu i przewidywaniu wzorców, często ważniejsze jest wykrywanie zmian w trendach. Zwykle wskazują one na ważną zmianę w mierzonym zjawisku, którą warto dodatkowo zbadać, aby ją wyjaśnić. W praktyce taką zmianą może być zmiana trendu cen akcji spółki na spadkowy po długich wzrostach. Rysunek 1.18 przedstawia przykładową zmianę trendu. Tu trend liniowy załamuje się po punkcie **x=50**.

Kup książkę Poleć książkę



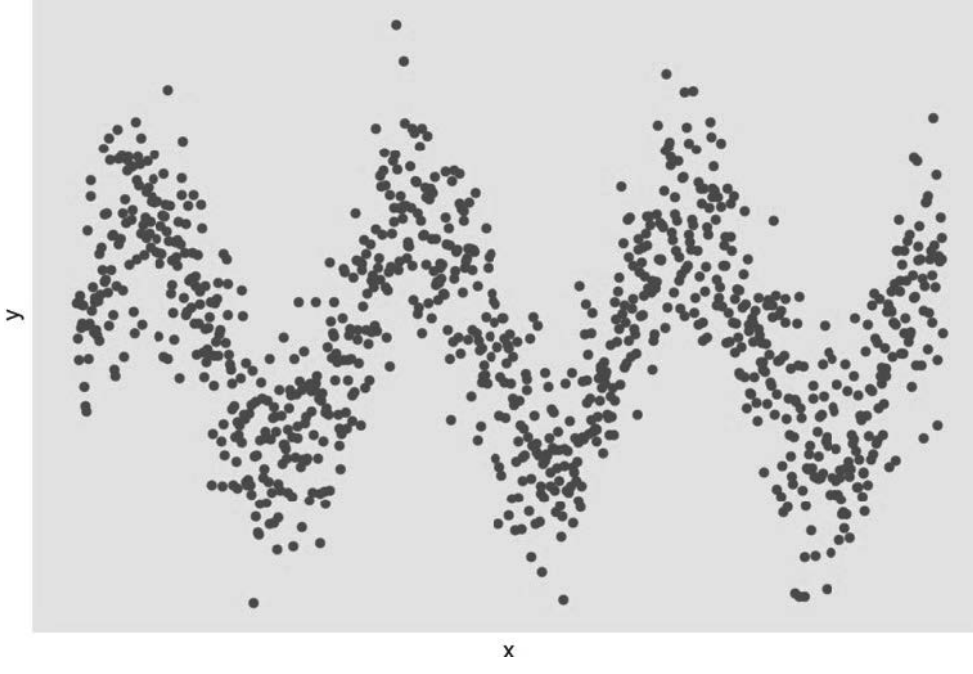
Rysunek 1.17. Inne często spotykane trendy



Rysunek 1.18. Przykład zmiany trendu

Kup książkę Poleć książkę

Innym wzorcem, na który często zwraca się uwagę, jest cykliczność, czyli powtarzające się wzorce w danych. Takie wzorce mogą wskazywać na to, że dwie zmienne zmieniają się cyklicznie, co może być przydatne w prognozowaniu. Rysunek 1.19 pokazuje przykład zmian cyklicznych.



Rysunek 1.19. Przykład zmian cyklicznych

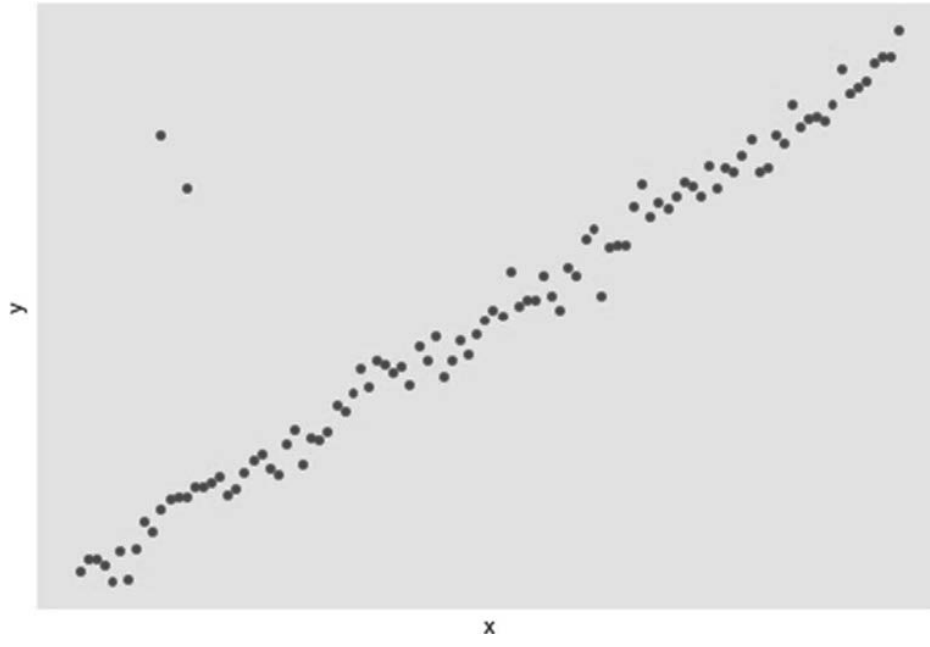
Wykresy punktowe umożliwiają też wykrywanie wartości odstających. Gdy większość punktów na wykresie znajduje się w określonym obszarze, ale niektóre są od niego znacznie oddalone, może to wskazywać, że te odległe punkty są wartościami odstającymi dla dwóch analizowanych zmiennych. W trakcie dalszych analiz dwuczynnikowych czasem warto pominąć takie punkty, aby ograniczyć szum w danych i uzyskać lepsze wnioski. Na rysunku 1.20 widoczne są punkty, które można uznać za wartości odstające.

Techniki bazujące na wykresach punktowych umożliwiają profesjonalnym analitykom danych zrozumienie ogólnych trendów w danych i wykonanie pierwszych kroków na drodze do przekształcenia danych w informacje.

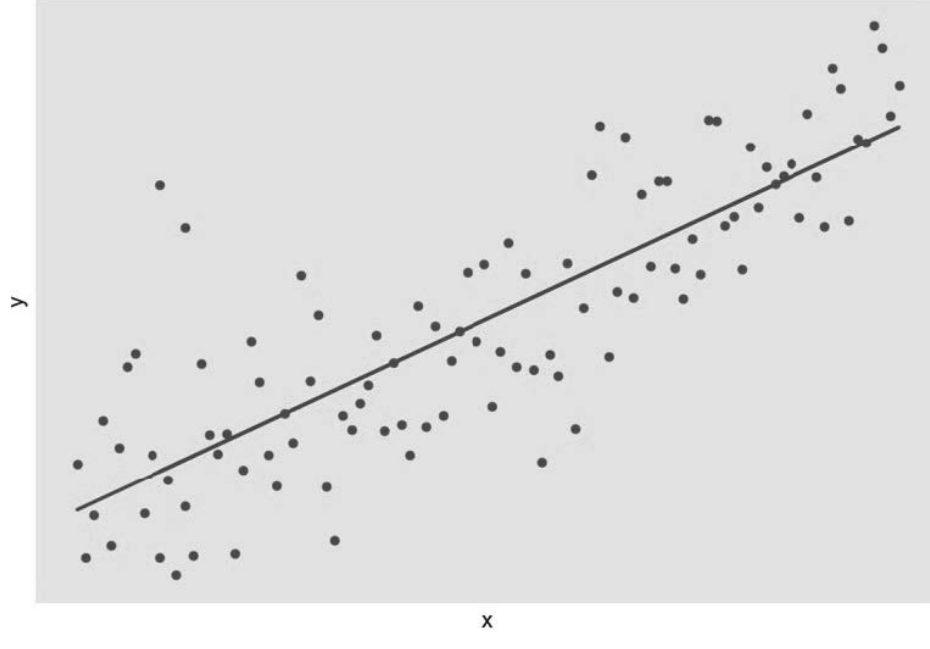
Współczynnik korelacji Pearsona

Jednym z najczęściej występujących trendów w trakcie analizy danych dwuczynnikowych jest trend liniowy. Często się zdarza jednak, że niektóre trendy liniowe pasują do danych lepiej, a inne gorzej. Na rysunkach 1.21 i 1.22 znajdziesz przykładowe wykresy punktowe z linią najlepszego dopasowania. Ta linia jest obliczana metodą **najmniejszych kwadratów**. Choć jej omawianie wykracza poza zakres tej książki, wiedza o tym, jak dobrze dane dwuczynnikowe pasują do trendu liniowego, pomaga zrozumieć zależność między dwiema zmiennymi.

Kup książkę Poleć książkę



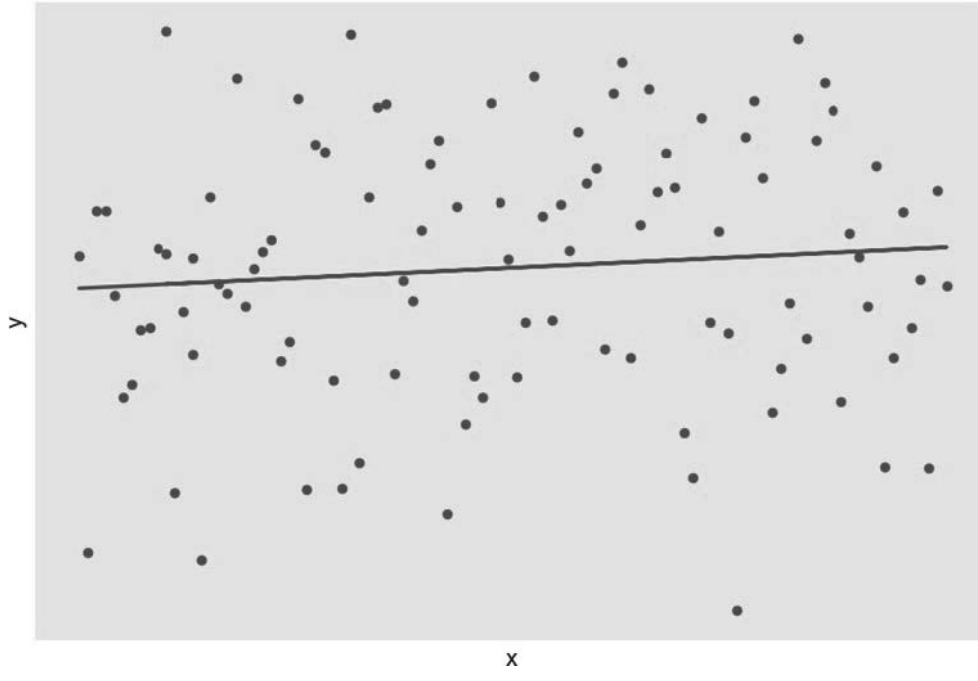
Rysunek 1.20. Wykres punktowy z dwiema wartościami odstającymi



Rysunek 1.21. Wykres punktowy z wyraźnym trendem liniowym

Kup książkę Poleć książkę

Kolejny rysunek ilustruje wykres punktowy ze słabym trendem liniowym.



Rysunek 1.22. Wykres punktowy ze słabym trendem liniowym

> Więcej o metodzie najmniejszych kwadratów dowiesz się z podręcznika do statystyki, na przykład z książki *Statistics* Davida Freedmana, Roberta Pisaniego i Rogera Purvesa.

Jedną z metod ilościowego reprezentowania korelacji liniowej jest użycie współczynnika korelacji Pearsona. Ten współczynnik, często zapisywany za pomocą litery r, to liczba z przedziału od –1 do 1 oznaczająca, jak dobrze wykres punktowy pasuje do trendu liniowego. Do obliczania współczynnika korelacji Pearsona (r) służy wzór z rysunku 1.23.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Rysunek 1.23. Wzór na obliczanie współczynnika korelacji Pearsona

Te obliczenia są dość skomplikowane, dlatego prześledź przykład, aby przekształcić wzór na konkretne kroki.

Kup książkę Poleć książkę

Ćwiczenie 1.05 — obliczanie współczynnika korelacji Pearsona dla dwóch zmiennych

W tym ćwiczeniu obliczysz współczynnik korelacji Pearsona dla relacji między godzinami przepracowanymi w tygodniu (`Hours Worked Per Week`) a sprzedażą tygodniową (`Sales Per Week ($)`). Na rysunku 1.24 widoczne są dane na temat 10 sprzedawców z salonu samochodowego firmy ZoomZoom z Houston, między innymi przychody z analizowanego tygodnia.

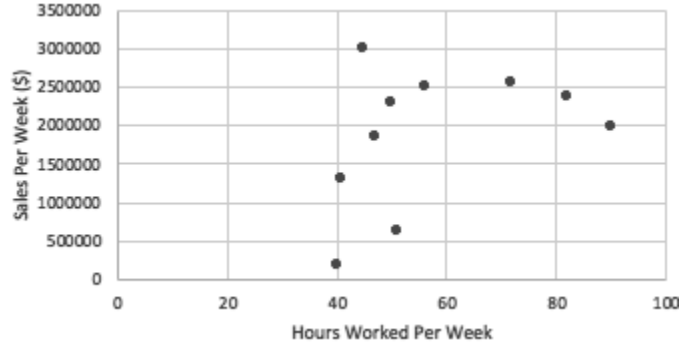
Hours Worked Per Week	Sales Per Week ($)
40	179480,58
56	2495037,37
50	2285369,51
82	2367896,33
41	1309745,16
51	623013,69
45	2989943,37
90	1970316,24
47	1845840,39
72	2553231,33

Rysunek 1.24. Dane 10 sprzedawców z salonu firmy ZoomZoom

Zbiór danych *salesman.csv* potrzebny w tym ćwiczeniu możesz bezpośrednio pobrać z serwisu GitHub. Oto odsyłacz do katalogu *Datasets* — *https://packt.live/2B1apb3*.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Najpierw utwórz wykres dwóch zmiennych w Excelu, używając danych z tego scenariusza. Pomoże Ci to na ogólnym poziomie ocenić, jakiego współczynnika korelacji Pearsona możesz oczekiwać.



Rysunek 1.25. Wykres punktowy godzin przepracowanych w tygodniu i wartości tygodniowej sprzedaży

Nie widać tu silnej liniowej zależności, ale wygląda na to, że tygodniowa sprzedaż rośnie wraz z liczbą przepracowanych godzin.

Kup książkę Poleć książkę

2. Teraz oblicz średnią każdej zmiennej. Powinieneś uzyskać 57,40 dla zmiennej `Hours Worked Per Week` i 1 861 987,3 dla zmiennej `Sales Per Week ($)`. Jeśli nie masz pewności, jak obliczyć średnią, zajrzyj do punktu „Tendencja centralna".
3. Teraz dla każdego wiersza oblicz cztery wartości: różnicę między wartością a średnią dla obu zmiennych oraz kwadrat tej różnicy dla obu zmiennych. Następnie oblicz iloczyn różnic. Powinieneś uzyskać tabelę wartości widoczną na rysunku 1.26.

Hours Worked Per Week	Sales Per Week ($)	x-mean(x)	(x-mean(x))^2	y-mean(y)	(y-mean(y))^2	[x-mean(x)][y-mean(y)]
40	179480,58	-17,4	302,76	-1682506,8	2830829189251,47	29275618,62
56	2495037,37	-1,4	1,96	633049,973	400752268315,30	-886269,96
50	2285369,51	-7,4	54,76	423382,113	179252413608,34	-3133027,64
82	2367896,33	24,6	605,16	505908,933	255943848489,20	12445359,75
41	1309745,16	-16,4	268,96	-552242,24	304971488326,76	9056772,69
51	623013,69	-6,4	40,96	-1238973,7	1535055846637,32	7929431,72
45	2989943,37	-12,4	153,76	1127955,97	1272284677026,38	-13986654,07
90	1970316,24	32,6	1062,76	108328,843	11735138225,72	3531520,28
47	1845840,39	-10,4	108,16	-16147,007	260725835,06	167928,87
72	2553231,33	14,6	213,16	691243,933	477818174909,31	10092161,42

Rysunek 1.26. Obliczenia współczynnika korelacji Pearsona

4. Oblicz sumy kwadratów i sumę iloczynów różnic. Powinieneś otrzymać 2812,40 dla zmiennej `Hours Worked Per Week` (x), 7 268 904 222 394,36 dla zmiennej `Sales Per Week ($)` (y) i 54 492 841,32 dla iloczynu różnic.
5. Oblicz pierwiastki kwadratowe sum różnic. Powinieneś uzyskać 53,03 dla zmiennej `Hours Worked Per Week` (x) i 2 696 090,54 dla zmiennej `Sales Per Week ($)` (y).
6. Podstaw te wartości do wzoru z rysunku 1.27. Otrzymasz wynik *0,38*. Obliczenia są pokazane na rysunku 1.27.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{54492841{,}32}{(53{,}03) * (2696090{,}54)} \approx 0{,}38$$

Rysunek 1.27. Gotowe obliczenia współczynnika korelacji Pearsona

W tym ćwiczeniu zobaczyłeś, jak obliczyć współczynnik korelacji Pearsona dwóch zmiennych. Po zastosowaniu wzoru otrzymałeś końcowy wynik 0,38.

Interpretowanie i analizowanie współczynnika korelacji

Ręczne obliczanie współczynnika korelacji może być bardzo skomplikowane. Zwykle lepiej jest obliczać go z użyciem komputera. W rozdziale 2., „Przygotowywanie danych za pomocą SQL-a", zobaczysz, że współczynnik korelacji Pearsona można obliczyć za pomocą SQL-a.

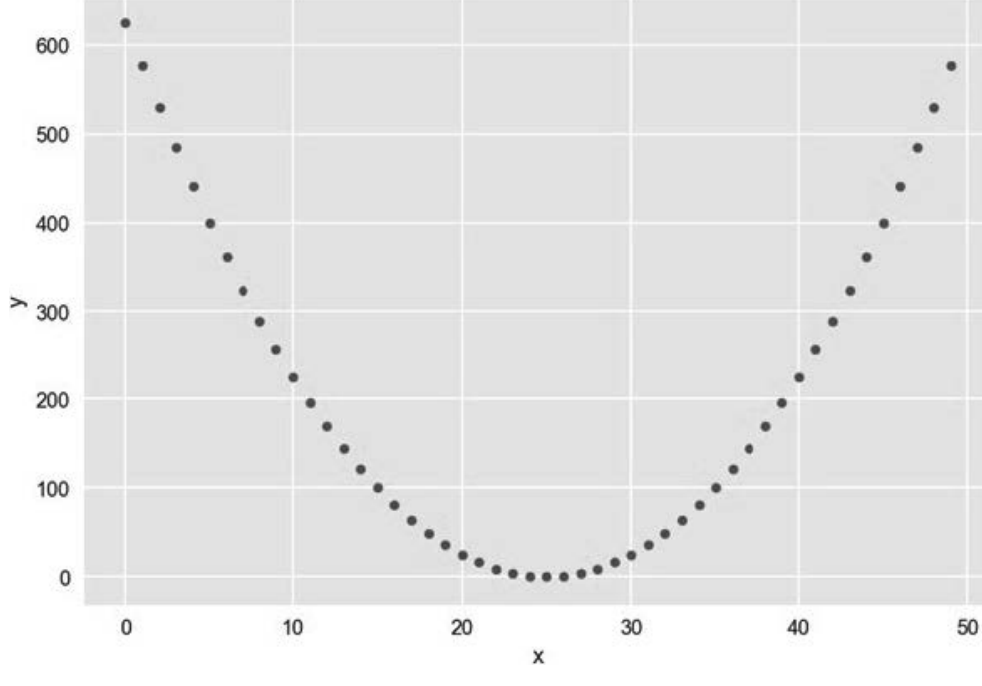
Aby zinterpretować współczynnik korelacji Pearsona, porównaj uzyskaną wartość z tabelą z rysunku 1.28. Im wynik jest bliższy zeru, tym korelacja jest słabsza. Im wyższa wartość bezwzględna współczynnika korelacji Pearsona, tym bardziej prawdopodobne jest, że punkty pasują do linii prostej.

Kup książkę Poleć książkę

Wartość korelacji	Interpretacja
–1,0 <= r <= –0,7	Bardzo mocna korelacja ujemna
–0,7 <= r <= –0,4	Mocna korelacja ujemna
–0,4 <= r <= –0,2	Umiarkowana korelacja ujemna
–0,2 <= r <= 0,2	Słaba lub nieistniejąca korelacja
0,2 <= r <= 0,4	Umiarkowana korelacja dodatnia
0,4 <= r <= 0,7	Mocna korelacja dodatnia
0,7 <= r <= 1,0	Bardzo mocna korelacja dodatnia

Rysunek 1.28. Interpretowanie współczynnika korelacji Pearsona

W trakcie analizowania współczynnika korelacji trzeba uwzględnić kilka kwestii. Pierwsza z nich dotyczy tego, że współczynnik korelacji mierzy, jak dobrze dwie zmienne pasują do trendu liniowego. Dwie zmienne mogą być ściśle powiązane, ale mieć stosunkowo niski współczynnik korelacji Pearsona. Przyjrzyj się na przykład punktom z rysunku 1.29. Jeśli obliczysz współczynnik dla tych dwóch zmiennych, otrzymasz wynik –0,08. Jednak krzywa wskazuje na bardzo silną zależność kwadratową. Dlatego gdy sprawdzasz współczynniki korelacji danych dwuczynnikowych, pamiętaj, że zależność między dwiema zmiennymi może być nieliniowa.

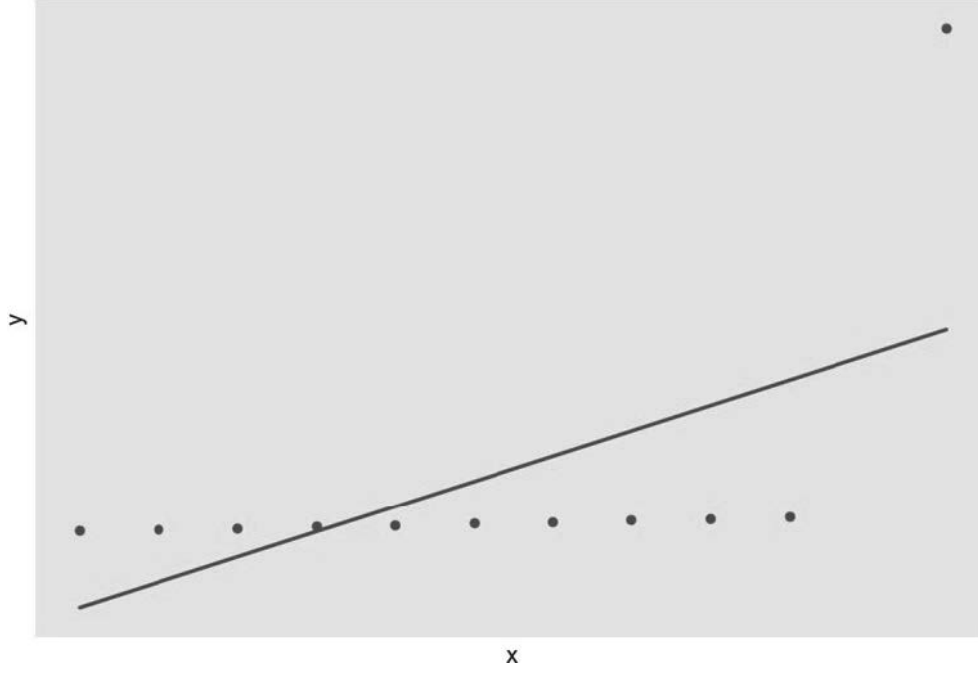


Rysunek 1.29. Silna nieliniowa zależność o niskim współczynniku korelacji

Kup książkę Poleć książkę

Innym ważnym aspektem jest liczba punktów używanych do obliczania korelacji. Wystarczą dwa punkty do zdefiniowania linii prostej. Dlatego mniejsza liczba punktów może skutkować otrzymaniem wysokiego współczynnika korelacji, który jednak nie zawsze zostanie utrzymany po dodaniu większej liczby danych. Zgodnie z ogólną regułą współczynniki korelacji obliczone dla mniej niż 30 punktów danych nie są wiarygodne. Do obliczania korelacji należy używać jak największej liczby dobrych punktów danych.

Zwróć uwagę na wyrażenie „dobre punkty danych". Jednym z powtarzających się motywów w tym rozdziale jest negatywny wpływ wartości odstających na różne statystyki. W danych dwuczynnikowych wartości odstające mogą wpływać na współczynnik korelacji. Przyjrzyj się wykresowi z rysunku 1.30. Widocznych jest tam 11 punktów, z których jeden to wartość odstająca. Powoduje on, że współczynnik korelacji Pearsona (*r*) dla tych danych spada do 0,59. Jednak bez tego punktu współczynnik jest równy 1,0. Dlatego należy starannie usunąć wartości odstające, zwłaszcza wtedy, gdy ich liczba jest mała.



Rysunek 1.30. Obliczanie r dla wykresu punktowego z wartością odstającą

Ważnym problemem związanym z obliczaniem korelacji jest błędne przyjmowanie, że korelacja oznacza związek przyczynowo-skutkowy. Występowanie wysokiej korelacji między *x* i *y* nie oznacza, że *x* powoduje *y*. Przyjrzyj się zależności między liczbą przepracowanych godzin a wartością sprzedanych dodatków. Przyjmij, że po dodaniu punktów danych okazuje się, że korelacja między tymi dwiema zmiennymi wynosi 0,5. Wielu początkujących analityków danych i doświadczonych menedżerów założy, że większa liczba przepracowanych godzin skutkuje wyższą sprzedażą, po czym zacznie zmuszać sprzedawców do nieustannej pracy.

Kup książkę Poleć książkę

Wprawdzie możliwe jest, że większa liczba godzin pracy skutkuje wyższą sprzedażą, jednak wysoki współczynnik korelacji nie wystarczy jako dowód tej tezy.

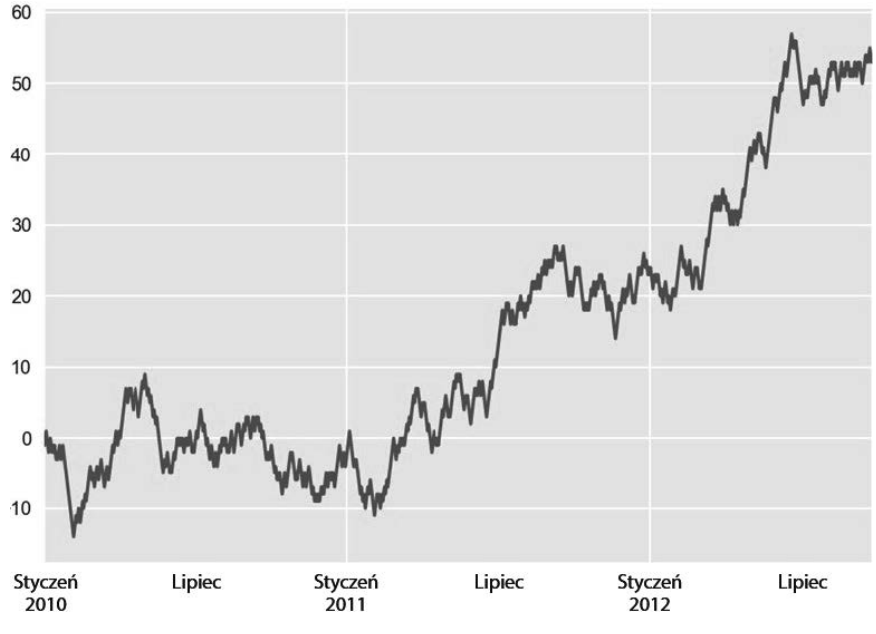
Możliwe nawet, że związek przyczynowy działa w drugą stronę: większa liczba transakcji wymaga więcej pracy papierkowej, co przekłada się na większą liczbę godzin w biurze. W tym scenariuszu większa liczba godzin nie powoduje wyższej sprzedaży.

Jeszcze inna możliwość to występowanie trzeciej zmiennej wpływającej na zależność między dwiema zmiennymi. Możliwe, że doświadczeni sprzedawcy pracują więcej godzin i lepiej radzą sobie ze sprzedażą. Dlatego prawdziwym powodem wyższej sprzedaży jest doświadczenie, z czego wynika zalecenie zatrudnienia większej grupy doświadczonych sprzedawców.

Profesjonalny analityk powinien unikać pułapek takich jak mylenie korelacji ze związkiem przyczynowym. Musisz krytycznie zastanowić się nad wszystkimi możliwościami, jakie wpływają na uzyskane wyniki.

Dane w postaci szeregów czasowych

Jednym z najważniejszych rodzajów analiz dwuczynnikowych jest analiza szeregów czasowych. **Szereg czasowy** reprezentuje relację dwuczynnikową, w której na osi *x* przedstawiony jest czas. Przykład szeregu czasowego jest pokazany na rysunku 1.31. Widoczny jest tam szereg czasowy obejmujący okres od stycznia 2010 roku do końca 2012 roku. Choć początkowo nie jest to oczywiste, daty i czas mają charakter ilościowy. Zrozumienie zmian zachodzących w czasie jest jednym z najważniejszych rodzajów analiz przeprowadzanych w firmach i zapewnia cenne informacje na temat kontekstu prowadzenia działalności.



Rysunek 1.31. Przykładowy szereg czasowy

Kup książkę Poleć książkę

Wszystkie wzorce opisane w poprzednim podrozdziale występują także w szeregach czasowych. Szeregi czasowe są ważne w firmach, ponieważ mogą wskazywać na czas wystąpienia zmian. Punkty w czasie mogą pomóc w ustaleniu przyczyn tych zmian.

Teraz przyjrzyj się niewielkiemu zbiorowi danych. Posłuży on do pokazania, jak przeprowadzać proste analizy statystyczne.

Zadanie 1.02 — eksplorowanie danych sprzedażowych z salonu samochodowego

W tym zadaniu przyjrzysz się dokładnie zbiorowi danych, wykorzystując statystykę. Wyobraź sobie, że jesteś analitykiem w ZoomZoom, firmie specjalizującej się w sprzedaży samochodów elektrycznych, i przeprowadzasz wysokopoziomowe analizy rocznej sprzedaży w salonach z całego kraju. Dane znajdują się w pliku *.csv*.

1. Otwórz dokument *dealerships.csv* w arkuszu kalkulacyjnym lub edytorze tekstu. Plik ten znajdziesz w katalogu *Datasets* w repozytorium w serwisie GitHub.
2. Przygotuj rozkład liczby kobiet zatrudnionych w poszczególnych salonach.
3. Ustal średnią i medianę dla rocznej sprzedaży salonu.
4. Oblicz odchylenie standardowe sprzedaży.
5. Czy dane z któregoś salonu są odstające? Wyjaśnij, dlaczego tak uważasz.
6. Oblicz kwantyle na podstawie rocznej sprzedaży.
7. Oblicz współczynnik korelacji rocznej sprzedaży z liczbą zatrudnionych kobiet i zinterpretuj wyniki.

Rozwiązanie tego zadania znajdziesz w „Dodatku”.

8. W tym zadaniu dane są kompletne. Co jednak zrobić, jeśli jest inaczej? Jak radzić sobie z brakiem danych? Następny punkt pomoże Ci zrozumieć, co robić w takich sytuacjach.

Praca z niepełnymi danymi

We wszystkich dotychczasowych przykładach zbiory danych były świetnie oczyszczone. Jednak w praktyce zbiory danych prawie nigdy nie są tak prawidłowe. Jednym z wielu problemów, z jakimi trzeba sobie radzić w trakcie pracy z danymi, są brakujące wartości. Szczegóły przygotowywania danych omówiliśmy w rozdziale 2., „Przygotowywanie danych za pomocą SQL-a”. Tu omawiamy kilka strategii, które możesz zastosować do radzenia sobie z niepełnymi danymi. Oto kilka możliwości:

- **Usuwanie wierszy** — jeśli danych brakuje w bardzo niewielkiej części wierszy (w mniej niż 5% zbioru danych), najprostszym rozwiązaniem może być usunięcie niepełnych punktów danych. Nie powinno to mieć istotnego wpływu na wyniki.

Kup książkę Poleć książkę

- **Wykorzystanie średniej, mediany lub wartości modalnej** — jeżeli wartość zmiennej jest nieobecna w od 5% do 25% punktów danych, możesz obliczyć średnią, medianę lub wartość modalną dla danej kolumny i uzupełnić luki otrzymanym wynikiem. Może to wprowadzać niewielką tendencyjność w obliczeniach, ale pozwala przeprowadzić więcej analiz bez usuwania cennych danych.
- **Wykorzystanie regresji** — jeśli jest to możliwe, przygotuj i zastosuj model do oszacowania brakujących wartości. Może to przekraczać możliwości większości analityków danych, jeśli jednak pracujesz ze specjalistą od data science, rozwiązanie to może być wykonalne.
- **Usuwanie zmiennej** — nie da się analizować nieistniejących danych. Jeśli danych jest niewiele, a w większości obserwacji brakuje wartości określonych zmiennej, lepszym rozwiązaniem może być usunięcie tej zmiennej niż przyjmowanie zbyt wielu założeń i dochodzenie do błędnych wniosków.

Przekonasz się, że analiza danych często jest bardziej sztuką niż nauką. Dotyczy to między innymi pracy z niekompletnymi danymi. Gdy nabierzesz doświadczenia, będziesz znać kombinacje strategii, które sprawdzają się w różnych scenariuszach.

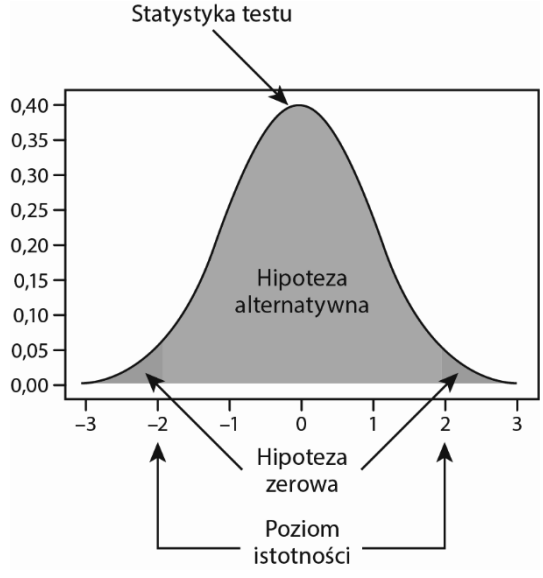
Testy istotności statystycznej

Następną czynnością przydatną w trakcie analiz danych jest sprawdzanie istotności statystycznej. Analitycy często porównują cechy statystyczne dwóch grup (lub tej samej grupy przed wprowadzeniem zmiany i po niej). Różnice między grupami mogą oczywiście wynikać z przypadku.

Ta technika jest stosowana na przykład w marketingowych testach A/B. Firmy często testują dwa rodzaje stron wejściowych produktu i mierzą **współczynnik klikalności** (ang. *click-through rate* — CTR). Może się okazać, że współczynnik klikalności dla wersji A strony wejściowej wynosi 10%, a dla wersji B — 11%. Czy to oznacza, że wersja B jest o 10% lepsza od wersji A? A może różnica wynika wyłącznie z przypadku i każdego dnia może być inna? Testy statystyczne pomagają odpowiedzieć na takie pytania.

W testach istotności statystycznej trzeba uwzględnić kilka ważnych elementów (rysunek 1.32). Przede wszystkim jest to **statystyka testu**. Może to być stosunek, średnia, różnica między grupami lub rozkład. Następnym niezbędnym elementem jest **hipoteza zerowa**, która zakłada, że zaobserwowane wyniki uzyskano przypadkowo. Potrzebna jest też **hipoteza alternatywna**, zgodnie z którą uzyskane wyniki nie są dziełem przypadku. Należy też ustalić **poziom istotności**, czyli wartość, jaką musi mieć statystyka testu, aby można było uznać, że hipoteza zerowa nie wyjaśnia różnic między grupami. Te cztery elementy występują we wszystkich testach istotności statystycznej, a różnice między testami wynikają ze sposobów obliczania tych komponentów.

Kup książkę Poleć książkę



Rysunek 1.32. Elementy testów istotności statystycznej

Często używane testy istotności statystycznej

Oto kilka często stosowanych testów istotności statystycznej:

- **Test Z dla dwóch próbek** — jest to test określający, czy średnie dwóch próbek różnią się od siebie. Ten test bazuje na założeniu, że obie próbki pochodzą z rozkładu normalnego o znanym odchyleniu standardowym dla populacji.
- **Test T dla dwóch próbek** — jest to test określający, czy średnie dwóch próbek różnią się od siebie. Stosuje się go, gdy albo próbki są zbyt małe (poniżej 30 punktów danych na próbkę), albo nieznane jest odchylenie standardowe dla populacji. Także tu przyjmuje się, że obie próbki pochodzą z populacji o rozkładzie normalnym.
- **Test zgodności chi-kwadrat (inaczej test Pearsona)** — ten test określa, czy rozkład punktów danych między kategorie różni się od oczekiwanego przypadkowego rozkładu. Służy przede wszystkim do sprawdzania, czy proporcje w testach (na przykład w testach A/B) różnią się od proporcji, jakie można uzyskać przypadkowo.

> Więcej o istotności statystycznej dowiesz się z podręcznika do statystyki, na przykład z książki *Statistics* Davida Freedmana, Roberta Pisaniego i Rogera Purvesa.

W następnym podrozdziale poznasz podstawy relacyjnych baz danych i SQL-a. Dalej omawiamy typy danych, polecenia i kwerendy.

Kup książkę Poleć książkę

Relacyjne bazy danych i SQL

Relacyjna baza danych to baza, w której używany jest **relacyjny model** danych. Model relacyjny został wymyślony w 1970 roku przez Edgara F. Codda i cechuje się tym, że dane są uporządkowane za pomocą relacji jako zbiory krotek. Każda krotka obejmuje zestaw atrybutów, które ją opisują. Wyobraź sobie relację, w której każda krotka reprezentuje klienta. Krotki mają tu atrybuty opisujące jednego klienta — na przykład jego nazwisko, imię i wiek w formacie (Sawicki, Jan, 27). Do unikatowego identyfikowania krotki w relacji służy **klucz** w postaci jednego atrybutu lub zestawu atrybutów. W modelu relacyjnym możliwe jest wykonywanie operacji logicznych na relacjach.

W bazie relacyjnej relacje mają zwykle postać tabel, podobnych jak w arkuszach kalkulacyjnych Excela. Każdy wiersz tabeli to krotka, a atrybuty są reprezentowane w kolumnach tabeli. Choć technicznie nie jest to wymagane, większość tabel w relacyjnych bazach danych ma kolumnę nazywaną **kluczem głównym**, która w unikatowy sposób identyfikuje wiersz bazy. Każda kolumna ma też **typ danych**, opisujący rodzaj danych przechowywanych w tej kolumnie. Tabele są zwykle łączone w kolekcje w bazach nazywane **schematami**. Wczytywanie tabel odbywa się w ramach **procesu ETL** (od ang. *extract*, *transform*, *load*, czyli pobieranie, przekształcanie i wczytywanie).

W kwerendach tabele przeważnie podaje się w formacie `[schemat].[tabela]`. Na przykład tabelę `products` ze schematu `analytics` zwykle będziesz podawać jako `analytics.products`. Istnieje też specjalny schemat o nazwie **`public`**. Jest to schemat domyślny, używany, gdy programista nie poda bezpośrednio schematu. Na przykład zapisy `public.products` i `products` oznaczają tę samą tabelę.

Oprogramowanie służące do zarządzania relacyjnymi bazami danych na komputerze to **system zarządzania relacyjnymi bazami danych** (**SZRBD**). SQL to język stosowany przez użytkowników SZRBD do dostępu do relacyjnych baz danych i interakcji z nimi.

> Prawie wszystkie relacyjne bazy danych oparte na SQL-u na jakimś poziomie są niezgodne z modelem relacyjnym. Na przykład nie każda tabela ma klucz. Ponadto model relacyjny nie dopuszcza powtarzających się wierszy, ale w relacyjnych bazach mogą one występować. Te rozbieżności są jednak niewielkie i dla zdecydowanej większości Czytelników tej książki nie mają żadnego znaczenia.

Wady i zalety baz SQL-owych

Od czasu wprowadzenia bazy danych Oracle w 1979 roku SQL stał się standardowym językiem do pracy z danymi w prawie wszystkich zastosowaniach informatycznych. Jest to uzasadnione. Bazy SQL-owe mają mnóstwo zalet, dlatego w wielu zastosowaniach wybiera się je prawie automatycznie. Oto te zalety:

Kup książkę Poleć książkę

- **Intuicyjność** — relacje w postaci tabel to często używana struktura danych zrozumiała prawie dla każdego. Dlatego praca z bazami relacyjnymi i myślenie o nich są znacznie prostsze niż w innych modelach.
- **Wydajność** — dzięki technice normalizacji bazy relacyjne umożliwiają reprezentowanie danych bez niepotrzebnego ich duplikowania. Dlatego pozwalają reprezentować dużą ilość informacji za pomocą niewielkiej ilości pamięci. To ograniczone zużycie pamięci zmniejsza też koszty operacyjne, dlatego przetwarzanie dobrze zaprojektowanych baz relacyjnych jest szybkie.
- **Deklaratywność** — SQL jest językiem deklaratywnym, co oznacza, że gdy piszesz kod, wystarczy poinformować komputer, jakie dane są potrzebne, a SZRBD zadba o ustalenie, jak wykonać kod w SQL-u. Nigdy nie musisz martwić się o przekazywanie komputerowi, jak ma uzyskać dostęp do danych z tabeli i je pobrać.
- **Niezawodność** — większość popularnych SQL-owych baz danych jest zgodna z właściwościami ACID (ang. *atomicity*, *consistency*, *isolation*, *durability*, czyli atomowość, spójność, izolacja i trwałość), co gwarantuje poprawność danych nawet po awariach sprzętu.

Bazy SQL-owe mają też jednak kilka wad. Oto one:

- **Stosunkowo niska specyficzność** — choć SQL jest deklaratywny, jego możliwości często są ograniczone do mechanizmów, jakie już zostały w nim zaprogramowane. Mimo że większość popularnych SZRBD jest nieustannie rozbudowywana o nowe możliwości, trudno jest pracować ze strukturami danych i algorytmami, które nie są zaprojektowane w używanym SZRBD.
- **Ograniczona skalowalność** — bazy SQL-owe są wysoce niezawodne, ma to jednak swoje koszty. Gdy przechowywana ilość informacji się podwaja, podwaja się też koszt zasobów. W przypadku przechowywania bardzo dużych zbiorów informacji lepszym wyborem mogą być inne magazyny danych, na przykład bazy typu NoSQL.
- **Niedopasowanie impedancji** — chociaż tabele to bardzo intuicyjna struktura danych, nie zawsze są najlepszym formatem do reprezentowania obiektów w komputerze. Wynika to przede wszystkim z tego, że obiekty często mają atrybuty z relacjami wiele do wielu. Na przykład klient firmy może kupić wiele produktów, a każdy produkt może zostać zakupiony przez wielu klientów. W obiektach w komputerze produkty można łatwo zapisać jako atrybut `lista` w obiekcie `klient`. Jednak w znormalizowanej bazie danych produkty klienta zapewne trzeba będzie zapisać za pomocą trzech różnych tabel, z których każda będzie wymagać aktualizacji po nowym zakupie, zwrocie lub wycofaniu ze sprzedaży.

Kup książkę Poleć książkę

Podstawowe typy danych w SQL-u

Wcześniej wspomnieliśmy, że każda kolumna w tabeli ma typ danych. Tu opisujemy podstawowe typy danych.

Typy liczbowe

Liczbowe typy danych reprezentują liczby. Na rysunku 1.33 pokazane są wybrane spośród popularnych typów liczbowych.

Nazwa	Zajmowana pamięć	Opis	Zakres
`smallint`	2 bajty	Małe liczby całkowite	Od –32 768 do +32 767
`integer`	4 bajty	Typowy wybór dla liczb całkowitych	Od –2 147 483 648 do +2 147 483 647
`bigint`	8 bajtów	Duże liczby całkowite	Od –9 223 372 036 854 775 808 do +9 223 372 036 854 775 807
`decimal`	Zmienna	Precyzja określona przez użytkownika, dokładne wartości	Do 131 072 cyfr przed przecinkiem; do 16 383 cyfr po przecinku
`numeric`	Zmienna	Precyzja określona przez użytkownika, dokładne wartości	Do 131 072 cyfr przed przecinkiem; do 16 383 cyfr po przecinku
`real`	4 bajty	Zmienna precyzja, niedokładne wartości	Precyzja do 6 cyfr po przecinku
`double`	8 bajtów	Zmienna precyzja, niedokładne wartości	Precyzja do 15 cyfr po przecinku
`smallserial`	2 bajty	Małe automatycznie zwiększane liczby całkowite	Od 1 do 32 767
`serial`	4 bajty	Automatycznie zwiększane liczby całkowite	Od 1 do 2 147 483 647
`bigserial`	8 bajtów	Duże automatycznie zwiększane liczby całkowite	Od 1 do 9 223 372 036 854 775 807

Rysunek 1.33. Podstawowe liczbowe typy danych

Typy znakowe

Znakowe typy danych przechowują informacje tekstowe. Na rysunku 1.34 znajdziesz opis znakowych typów danych.

Kup książkę Poleć książkę

Nazwa	Opis
`character varying(n)`, `varchar(n)`	Zmienna ograniczona długość
`character(n)`, `char(n)`	Stała długość, dopełnianie spacjami
`text`	Zmienna nieograniczona długość

Rysunek 1.34. Podstawowe znakowe typy danych

W systemie PostgreSQL i wielu innych bazach SQL-owych na zapleczu dla wszystkich znakowych typów danych używana jest ta sama struktura danych. Większość programistów nie stosuje obecnie typu `char(n)`.

Typ logiczny

Logiczny typ danych służy do reprezentowania wartości `True` lub `False`. W tabeli na rysunku 1.35 opisane są wartości traktowane jako logiczne w kwerendach z kolumną typu logicznego.

Wartość logiczna	Akceptowane wartości
`True`	`t`, `true`, `y`, `yes`, `on`, `1`
`False`	`f`, `false`, `n`, `no`, `off`, `0`

Rysunek 1.35. Akceptowane wartości logiczne

Choć akceptowane są wszystkie wymienione wartości, zgodnie z najlepszymi praktykami należy używać wartości `True` i `False`. W kolumnach logicznych dopuszczalne mogą być także wartości `NULL`.

Daty i godziny

Typy danych z datą i godziną służą do zapisywania informacji związanych z czasem, na przykład dat i godzin. Na rysunku 1.36 pokazane są przykładowe typy danych z datą i godziną.

Nazwa	Wielkość	Opis
`Timestamp without timezone`	8 bajtów	Data i czas bez strefy czasowej
`Timestamp with timezone`	8 bajtów	Data i czas ze strefą czasową
`date`	4 bajty	Data (bez czasu)
`Time without timezone`	8 bajtów	Czas (bez daty)
`Time with timezone`	12 bajtów	Sam czas ze strefą czasową
`interval`	16 bajtów	Przedział czasu

Rysunek 1.36. Popularne typy danych z datą i godziną

Ten typ danych omawiamy dokładniej w rozdziale 5., „Analizy z wykorzystaniem złożonych typów danych”.

Kup książkę Poleć książkę

Struktury danych — format JSON i tablice

Wiele wersji SQL-a udostępnia też struktury danych takie jak obiekty w formacie **JSON** (ang. *JavaScript Object Notation*) i tablice. Tablice to listy danych zwykle zapisywane jako elementy w nawiasie klamrowym. Oto przykładowa tablica: `['kot', 'pies', 'koń']`. Obiekt w formacie JSON to zbiór par klucz – wartość rozdzielonych przecinkami i umieszczonych w nawiasie klamrowym. Na przykład `{'imię': 'Borys', 'wiek': 27, 'miasto': 'Nowy Sącz'}` to poprawny obiekt w formacie JSON. Tego rodzaju struktury danych często występują w aplikacjach, a możliwość stosowania ich w bazach danych ułatwia wykonywanie wielu analiz.

Struktury danych opisujemy szczegółowo w rozdziale 5., „Analizy z wykorzystaniem złożonych typów danych".

Przyjrzyj się teraz podstawowym operacjom z użyciem SQL-a w SZRBD.

Wczytywanie tabel — kwerenda SELECT

Najczęściej wykonywaną operacją w bazach danych jest wczytywanie danych z bazy. Prawie zawsze używane jest do tego słowo kluczowe `SELECT`.

Podstawowa budowa i działanie kwerendy SELECT

Na ogólnym poziomie kwerendę można podzielić na pięć części:

- **Operacja** — pierwsza część kwerendy opisuje, jaka operacja zostanie wykonana. Tu używane jest słowo `SELECT`, po którym podawane są nazwy kolumn i funkcji.
- **Dane** — następną część kwerendy stanowią dane. Podaje się je jako słowo kluczowe `FROM`, po którym następuje nazwa tabeli (lub kilku tabel) wraz z zarezerwowanymi słowami kluczowymi określającymi, jakie dane należy uwzględnić na potrzeby wykonywania obliczeń, filtrowania i pobierania.
- **Warunek** — ta część kwerendy filtruje dane, aby uwzględniane były tylko te wiersze, które spełniają podany warunek, zwykle określony w klauzuli `WHERE`.
- **Grupowanie** — ten etap bazuje na specjalnej klauzuli, która pobiera wiersze ze źródła danych i łączy je ze sobą na podstawie klucza z klauzuli `GROUP BY`, a następnie oblicza wynik na podstawie wartości z wszystkich wierszy o tym samym kluczu. Ten krok jest opisany szczegółowo w rozdziale 3., „Agregacja i funkcje okna".
- **Przetwarzanie końcowe —** na tym etapie kwerendy wynikowe dane są pobierane i przetwarzane (sortowanie i uwzględnianie ograniczeń), często z użyciem słów kluczowych takich jak `ORDER BY` i `LIMIT`.

Kup książkę Poleć książkę

Oto etapy działania kwerendy SELECT:

1. Tworzenie źródła danych. Podaj jedną tabelę lub wskaż kilka tabel, które zostaną połączone w jedną dużą tabelę.
2. Filtrowanie tabeli z dużego źródła danych utworzonego w kroku 1.. W tym celu sprawdzane jest, które wiersze są zgodne z klauzulą WHERE.
3. Obliczanie wartości na podstawie kolumn ze źródła danych z kroku 1. Jeśli używana jest klauzula GROUP BY, wiersze są dzielone na grupy, po czym obliczane są statystyki zbiorcze dla każdej grupy. W przeciwnym razie zwracana jest kolumna lub wartość obliczona przez wykonanie funkcji na jednej kolumnie lub kilku kolumnach.
4. Pobranie zwróconych wierszy i uporządkowanie ich zgodnie z kwerendą.

Aby przeanalizować te kroki, przyjrzyj się typowej kwerendzie i prześledź opisany proces:

```
SELECT
  first_name
FROM
  customers
WHERE
  state='AZ'
ORDER BY
  first_name;
```

Ta kwerenda działa w następujący sposób:

1. Punktem wyjścia jest tabela customers.
2. Ta tabela jest filtrowana; uwzględniane są tylko te wiersze, w których kolumna state ma wartość 'AZ'.
3. Z przefiltrowanej tabeli pobierana jest kolumna first_name.
4. Wartości z kolumny first_name są sortowane alfabetycznie.

Pokazaliśmy tu, jak rozbić kwerendę na serię kroków do wykonania przez bazę danych. Teraz poznasz słowa kluczowe używane w kwerendach SELECT i wzorce takich kwerend.

Podstawowe słowa kluczowe w kwerendach SELECT

W trakcie pisania kwerendy SELECT możesz stosować wiele słów kluczowych. Na początek poznaj słowa kluczowe SELECT i FROM.

Instrukcje SELECT i FROM

Najbardziej podstawowa wersja kwerendy SELECT ma postać SELECT…FROM <nazwa_tabeli>;. Ta kwerenda pozwala pobrać dane z jednej tabeli. Na przykład jeśli chcesz pobrać wszystkie dane z tabeli products z przykładowej bazy, użyj następującej kwerendy:

```
SELECT
  *
FROM
  products;
```


Kup książkę Poleć książkę

Ta kwerenda pobiera wszystkie dane z bazy. Użyty tu symbol * to skrót pozwalający pobrać z bazy wszystkie kolumny. Średnik (;) informuje komputer, że dotarł do końca kwerendy, podobnie jak kropka kończy zwykłe zdania. Należy zauważyć, że wiersze są tu zwracane w nieokreślonej kolejności. Jeśli chcesz pobrać w kwerendzie tylko określone kolumny, zastąp gwiazdkę (*) nazwami potrzebnych kolumn. Kolumny należy rozdzielić przecinkami i podać w kolejności, w jakiej kwerenda ma zwracać dane. Na przykład jeśli chcesz otrzymać kolumnę `product_id`, po której następuje kolumna `model` (obie z tabeli `products`), użyj tej kwerendy:

```
SELECT product_id, model
FROM products;
```

Jeśli chcesz, aby najpierw zwracana była kolumna `model`, a następnie kolumna `product_id`, zapisz kwerendę tak:

```
SELECT model, product_id
FROM products;
```

W następnym punkcie poznasz klauzulę `WHERE`.

Klauzula WHERE

Klauzula `WHERE` pozwala dodać warunek ograniczający ilość zwracanych danych. Wszystkie wiersze zwracane przez kwerendę `SELECT` z klauzulą `WHERE` spełniają warunek z tej klauzuli. W kwerendzie `SELECT` klauzula `WHERE` zwykle jest umieszczana po klauzuli `FROM`.

Warunkiem w klauzuli `WHERE` zazwyczaj jest wyrażenie logiczne, które dla każdego wiersza przyjmuje wynik `True` lub `False`. Gdy używane są kolumny liczbowe, w wyrażeniu logicznym można stosować operatory równości, większości lub mniejszości, aby porównywać wartość z kolumny z podaną wartością.

Oto ilustrujący to przykład. Załóżmy, że chcesz sprawdzić nazwy modeli z 2014 roku z przykładowego zbioru danych. W tym celu możesz napisać następującą kwerendę:

```
SELECT
  model
FROM
  products
WHERE
  year=2014;
```

Z następnego punktu dowiesz się, jak stosować w kwerendach klauzule `AND` i `OR`.

Klauzule AND i OR

W ostatniej kwerendzie używany był tylko jeden warunek. Często analityka interesują dane spełniające kilka warunków. W tym celu można połączyć zestaw wyrażeń za pomocą klauzul `AND` i `OR`.

Kup książkę Poleć książkę

Przyjrzyj się ilustrującemu to przykładowi. Wyobraź sobie, że chcesz zwrócić modele, które nie tylko zostały zbudowane w 2014 roku, ale dodatkowo mają sugerowaną cenę detaliczną producenta niższą niż 1000 dolarów. Możesz napisać następującą kwerendę:

```
SELECT
  model
FROM
  products
WHERE
  year=2014
  AND msrp<=1000;
```

Teraz załóżmy, że chcesz zwrócić dowolne modele, które zostały wypuszczone w 2014 roku lub których typ produktu to `automobile`. Umożliwia to poniższa kwerenda:

```
SELECT
  model
FROM
  products
WHERE
  year=2014
  OR product_type='automobile';
```

Gdy używasz więcej niż jednej klauzuli `AND` lub `OR`, dodaj nawiasy, aby odpowiednio oddzielić i rozmieścić wyrażenia logiczne. To gwarantuje, że kwerenda będzie działać zgodnie z oczekiwaniami i że będzie czytelna. Jeśli na przykład chcesz pobrać wszystkie produkty z lat od 2014 do 2016, a także produkty typu `'scooter'`, możesz użyć następującej kwerendy:

```
SELECT
  *
FROM
  products
WHERE
  year>2014
  AND year<2016
  OR product_type='scooter';
```

Jednak aby klauzula `WHERE` była bardziej czytelna, lepiej jest użyć następującego zapisu:

```
SELECT
  *
FROM
  products
WHERE
  (year>2014 AND year<2016)
  OR product_type='scooter';
```

W następnym punkcie poznasz klauzule `IN` i `NOT IN`.

Klauzule IN i NOT IN

Wcześniej wspomnieliśmy, że w wyrażeniach logicznych można używać znaku równości, aby określić, że kolumna musi się równać określonej wartości. Co jednak zrobić, jeśli chcesz zwrócić wiersze, w których kolumna równa się jednej z grup wartości? Załóżmy, że interesują Cię wszystkie modele z lat 2014, 2016 i 2019. Możesz zapisać taką kwerendę tak:

Kup książkę Poleć książkę

```
SELECT
  model
FROM
  products
WHERE
  year = 2014
  OR year = 2016
  OR year = 2019;
```

Jednak ten kod jest długi i żmudny w pisaniu. Za pomocą klauzuli `IN` możesz zapisać go tak:

```
SELECT
  model
FROM
  products
WHERE
  year IN (2014, 2016, 2019);
```

Tę wersję znacznie łatwiej jest zapisać, a także zrozumieć.

Możesz też zastosować klauzulę `NOT IN`, aby zwrócić wszystkie wartości poza podanymi na liście. Jeśli chcesz otrzymać wszystkie towary, które nie zostały wyprodukowane w latach 2014, 2016 lub 2019, możesz użyć tej kwerendy:

```
SELECT
  model
FROM
  products
WHERE
  year NOT IN (2014, 2016, 2019);
```

W następnym punkcie dowiesz się, jak używać w kwerendach klauzuli `ORDER BY`.

Klauzula ORDER BY

Wcześniej wspomnieliśmy, że jeśli kwerenda SQL-owa nie otrzyma konkretnych instrukcji, uporządkuje wiersze tak, jak zostały one znalezione przez bazę danych. W wielu scenariuszach jest to akceptowalne. Jednak często pożądane jest zwracanie wierszy w określonym porządku. Wyobraź sobie, że chcesz pobrać wszystkie produkty uporządkowane od najstarszych do najnowszych według daty rozpoczęcia produkcji. W SQL-u umożliwia to klauzula `ORDER BY`:

```
SELECT
  model
FROM
  products
ORDER BY
  production_start_date;
```

Jeśli kolejność sortowania nie jest bezpośrednio określona, zwracane wiersze będą uporządkowane rosnąco. Oznacza to, że wiersze są porządkowane od najmniejszej do największej wartości z wybranej kolumny (lub ze zbioru kolumn). W przypadku tekstu uwzględniana jest

Kup książkę Poleć książkę

kolejność alfabetyczna. Aby jawnie zażądać kolejności rosnącej, użyj słowa kluczowego `ASC`. W poprzedniej kwerendzie można to zrobić tak:

```
SELECT
  model
FROM
  products
ORDER BY
  production_start_date ASC;
```

Jeśli chcesz pobrać dane w porządku malejącym, użyj słowa kluczowego `DESC`. Jeżeli zamierzasz wczytać modele uporządkowane od najnowszych do najstarszych, zastosuj następujący kod:

```
SELECT
  model
FROM
  products
ORDER BY
  production_start_date DESC;
```

Ponadto zamiast podawać nazwę kolumny, według której chcesz sortować dane, możesz podać numer tej kolumny w naturalnym układzie tabeli. Załóżmy, że chcesz zwrócić wszystkie modele z tabeli `products` uporządkowane według identyfikatorów produktów. Możesz zapisać kwerendę tak:

```
SELECT
  model
FROM
  products
ORDER BY
  product_id;
```

Ponieważ jednak `product_id` to pierwsza kolumna w tej tabeli, możesz też zastosować taki zapis:

```
SELECT
  model
FROM
  products
ORDER BY
  1;
```

Możesz także sortować dane według kilku kolumn, podając po klauzuli `ORDER BY` dodatkowe kolumny rozdzielone przecinkami. Załóżmy, że chcesz uporządkować wszystkie wiersze tabeli najpierw według roku produkcji modelu (od najnowszych do najstarszych), a następnie według sugerowanej ceny detalicznej (od najniższej do najwyższej). Możesz zapisać kod tak:

```
SELECT
  *
FROM
  products
ORDER BY
  year DESC,
  base_msrp ASC;
```


Kup książkę Poleć książkę

Na rysunku 1.37 widoczne są dane wyjściowe tego kodu.

product_id bigint	model text	year bigint	product_type text	base_msrp numeric	production_start_date timestamp without time zone	production_end_date timestamp without time zone
12	Lemon ...	2019	scooter	349.99	2019-02-04 00:00:00	[null]
11	Model ...	2019	automobile	95000.00	2019-02-04 00:00:00	[null]
8	Bat Limi...	2017	scooter	699.99	2017-02-15 00:00:00	[null]
9	Model E...	2017	automobile	35000.00	2017-02-15 00:00:00	[null]
10	Model ...	2017	automobile	85750.00	2017-02-15 00:00:00	[null]
7	Bat	2016	scooter	599.99	2016-10-10 00:00:00	[null]
6	Model S...	2015	automobile	65500.00	2015-04-15 00:00:00	2018-10-01 00:00:00
5	Blade	2014	scooter	699.99	2014-06-23 00:00:00	2015-01-27 00:00:00
4	Model ...	2014	automobile	115000.00	2014-06-23 00:00:00	2018-12-28 00:00:00
3	Lemon	2013	scooter	499.99	2013-05-01 00:00:00	2018-12-28 00:00:00
2	Lemon ...	2011	scooter	799.99	2011-01-03 00:00:00	2011-03-30 00:00:00

Rysunek 1.37. Porządkowanie według wielu kolumn za pomocą klauzuli ORDER BY

W następnym punkcie poznasz słowo kluczowe `LIMIT` z SQL-a.

Klauzula LIMIT

Większość tabel w bazach SQL-owych jest dość duża, dlatego nie trzeba zwracać wszystkich wierszy. Czasem potrzebnych jest tylko kilka pierwszych wierszy. W takim scenariuszu przydatne jest słowo kluczowe `LIMIT`. Wyobraź sobie, że chcesz pobrać tylko pięć pierwszych produktów firmy. Możesz to zrobić za pomocą następującej kwerendy:

```
SELECT
  model
FROM
  products
ORDER BY
  production_start_date
LIMIT
  5;
```

Na rysunku 1.38 pokazane są dane wyjściowe tego kodu.

model text
Lemon
Lemon Limited Edition
Lemon
Blade
Model Chi

Rysunek 1.38. Kwerenda z klauzulą LIMIT

Zgodnie z ogólną regułą zwykle warto używać słowa kluczowego `LIMIT` w tabelach i kwerendach, z których wcześniej nie korzystałeś.

Kup książkę Poleć książkę

Klauzule IS NULL i IS NOT NULL

Często w kolumnie brakuje niektórych wartości. Może to wynikać z wielu powodów. Możliwe, że dane nie zostały pobrane lub były niedostępne w momencie ich wprowadzania. Możliwe, że proces ETL nie pobrał danych i nie wczytał ich do kolumny. Możliwe też, że brak wartości reprezentuje określony stan wiersza i stanowi cenną informację. Niezależnie od powodu analityk często chce znaleźć wiersze, w których w określonej kolumnie brakuje wartości. W SQL-u brak wartości nieraz jest reprezentowany za pomocą `NULL`. Na przykład w tabeli `products` wartość `NULL` w kolumnie `production_end_date` oznacza, że produkt wciąż jest wytwarzany. Dlatego jeśli chcesz wyświetlić wszystkie nadal produkowane towary, możesz użyć następującej kwerendy:

```
SELECT
  *
FROM
  products
WHERE
  production_end_date IS NULL;
```

Na rysunku 1.39 pokazane są dane wyjściowe.

product_id bigint	model text	year bigint	product_type text	base_msrp text	production_start_date timestamp without time zone	production_end_date timestamp without time zone
7	Bat	2016	scooter	599.99	2016-10-10 00:00:00	[null]
8	Bat Limi...	2017	scooter	699.99	2017-02-15 00:00:00	[null]
9	Model E...	2017	automobile	35,000.00	2017-02-15 00:00:00	[null]
10	Model ...	2017	automobile	85,750.00	2017-02-15 00:00:00	[null]
11	Model ...	2019	automobile	95,000.00	2019-02-04 00:00:00	[null]
12	Lemon ...	2019	scooter	349.99	2019-02-04 00:00:00	[null]

Rysunek 1.39. Produkty z wartością NULL w kolumnie production_end_date

Jeśli interesują Cię tylko produkty, które już nie są wytwarzane, możesz użyć klauzuli `IS NOT NULL`, tak jak w tej kwerendzie:

```
SELECT *
FROM products
WHERE production_end_date IS NOT NULL;
```

Dane wyjściowe są pokazane na rysunku 1.40.

product_id bigint	model text	year bigint	product_type text	base_msrp text	production_start_date timestamp without time zone	production_end_date timestamp without time zone
1	Lemon	2010	scooter	399.99	2010-03-03 00:00:00	2012-06-08 00:00:00
2	Lemon ...	2011	scooter	799.99	2011-01-03 00:00:00	2011-03-30 00:00:00
3	Lemon	2013	scooter	499.99	2013-05-01 00:00:00	2018-12-28 00:00:00
4	Model ...	2014	automobile	115,000.00	2014-06-23 00:00:00	2018-12-28 00:00:00
5	Blade	2014	scooter	699.99	2014-06-23 00:00:00	2015-01-27 00:00:00
6	Model S...	2015	automobile	65,500.00	2015-04-15 00:00:00	2018-10-01 00:00:00

Rysunek 1.40. Produkty z kolumną production_end_date o wartości różnej od NULL

W następnym ćwiczeniu zobaczysz, jak wykorzystać nowo poznane słowa kluczowe.

Kup książkę Poleć książkę

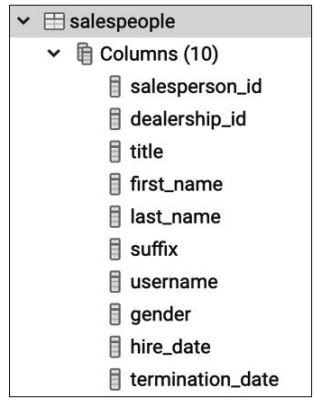
Ćwiczenie 1.06 — kwerenda SELECT z podstawowymi słowami kluczowymi dotycząca tabeli salespeople

W tym ćwiczeniu utworzysz kilka kwerend `SELECT`, używając podstawowych słów kluczowych. Wyobraź sobie, że po kilku dniach w nowej pracy wreszcie uzyskujesz dostęp do firmowej bazy danych. Dziś Twój szef prosi Cię, abyś pomógł menedżerowi ds. sprzedaży, który nie zna dobrze SQL-a. Menedżer chce uzyskać kilka różnych list z danymi o sprzedawcach. Najpierw przygotuj listę internetowych nazw użytkownika dziesięciu pierwszych kobiet pracujących na stanowisku sprzedawcy; dane uporządkuj od pierwszej do ostatniej zatrudnionej.

> We wszystkich kolejnych ćwiczeniach z tej książki używane jest narzędzie pgAdmin 4.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Otwórz wybranego klienta SQL-a i nawiąż połączenie z bazą `sqlda`.
2. Zbadaj schemat tabeli `salespeople` za pomocą listy rozwijanej. Zwróć uwagę na nazwy kolumn z rysunku 1.41.



Rysunek 1.41. Schemat tabeli salespeople

3. Wykonaj poniższą kwerendę, aby uzyskać nazwy użytkownika kobiet na stanowisku sprzedawcy posortowane według wartości `hire_date`; w klauzuli `LIMIT` użyj wartości 10:

```
SELECT
  username
FROM
  salespeople
WHERE
  gender= 'Female'
ORDER BY
  hire_date
  LIMIT 10;
```


Kup książkę Poleć książkę

Rysunek 1.42 przedstawia dane wyjściowe.

	username text
1	nlie2l
2	adufaire3r
3	bgrimoldby4q
4	jmedgewick...
5	bhain3y
6	kclyburn54
7	adobbing4g
8	skinner1h
9	alimon7j

Rysunek 1.42. Nazwy użytkownika kobiet pracujących jako sprzedawcy posortowane według daty zatrudnienia

W ten sposób otrzymujesz listę nazw użytkownika kobiet pracujących jako sprzedawca uporządkowanych od najwcześniej do najpóźniej zatrudnionych.

Kod źródłowy do tego punktu znajdziesz na stronie *https://packt.live/2B4qMUk*.

W tym ćwiczeniu użyłeś różnych podstawowych słów kluczowych w kwerendzie `SELECT`, aby pomóc menedżerowi ds. sprzedaży w pobraniu listy sprzedawców zgodnie z podanymi kryteriami.

Zadanie 1.03 — kwerenda SELECT z podstawowymi słowami kluczowymi dotycząca tabeli customers

Dział marketingu uznał, że chce przeprowadzić serię kampanii marketingowych, aby pomóc w zwiększeniu sprzedaży. W tym celu potrzebuje szczegółowych informacji o wszystkich klientach z Nowego Jorku. Oto kroki potrzebne do wykonania tego zadania:

1. Otwórz wybranego klienta SQL-a i nawiąż połączenie z bazą `sqlda`. Zbadaj schemat z tabeli `customers` za pomocą listy rozwijanej schematu.
2. Napisz kwerendę, która zwraca w porządku alfabetycznym e-maile wszystkich klientów firmy ZoomZoom ze stanu Floryda.
3. Napisz kwerendę, która zwraca imię, nazwisko i e-mail klientów firmy ZoomZoom z miasta Nowy Jork w stanie Nowy Jork. Dane mają być uporządkowane alfabetycznie najpierw według nazwiska, a następnie według imienia.

Kup książkę Poleć książkę

4. Napisz kwerendę, która zwraca dane wszystkich klientów wraz z numerem telefonu. Dane mają być uporządkowane według daty dodania klienta do bazy.

> Rozwiązanie tego zadania znajduje się w „Dodatku".

W tym zadaniu użyłeś różnych podstawowych słów kluczowych w kwerendzie `SELECT`, aby pomóc menedżerowi z działu marketingu w pobraniu danych potrzebnych do kampanii marketingowej.

Tworzenie tabel

Wiesz już, jak wczytywać dane z tabel. Teraz zobacz, jak tworzyć nowe tabele. Można to robić na dwa sposoby — tworząc puste tabele lub używając kwerend `SELECT`.

Tworzenie pustych tabel

Aby utworzyć nową pustą tabelę, należy użyć instrukcji `CREATE TABLE`. Ta instrukcja ma następującą strukturę:

```
CREATE TABLE {nazwa_tabeli} (
{nazwa_kolumny_1} {typ_danych_1} {ograniczenia_kolumny_1},
{nazwa_kolumny_2} {typ_danych_2} {ograniczenia_kolumny_2},
{nazwa_kolumny_3} {typ_danych_3} {ograniczenia_kolumny_3},
…
{nazwa_kolumny_ostatnia} {typ_danych_ostatni} {ograniczenia_kolumny_ostatnie},
);
```

Tu `{nazwa_tabeli}` to nazwa tabeli, `{nazwa_kolumny}` to nazwa kolumny, `{typ_danych}` to typ danych kolumny, a `{ograniczenia_kolumny}` to jedno lub kilka opcjonalnych słów kluczowych określających specjalne cechy kolumny. Przed omówieniem użycia kwerendy `CREATE TABLE` warto najpierw omówić ograniczenia kolumn.

Ograniczenia kolumn

Ograniczenia kolumn to słowa kluczowe określające specjalne cechy kolumn. Oto wybrane ważne ograniczenia:

- `NOT NULL` — gwarantuje, że żadna wartość w danej kolumnie nie będzie równa `NULL`.
- `UNIQUE` — gwarantuje, że każdy wiersz w kolumnę ma unikatową wartość (żadna wartość się nie powtarza).
- `PRIMARY KEY` — jest to specjalne ograniczenie wyznaczające kolumnę klucza głównego. Oznacza, że wartość w danej kolumnie jest unikatowa dla każdego wiersza i pomaga szybciej wyszukiwać wiersze. Tylko jedna kolumna tabeli może być kluczem głównym.

Kup książkę Poleć książkę

Załóżmy, że chcesz utworzyć tabelę `state_populations` z kolumnami ze skrótową nazwą i liczbą mieszkańców każdego stanu USA. Potrzebna kwerenda wygląda tak:

```
CREATE TABLE state_populations (
  state VARCHAR(2) PRIMARY KEY,
  population NUMERIC
);
```

Ta kwerenda daje następujący wynik:

```
Query returned successfully in 122 msec.
```

> Czasem po uruchomieniu kwerendy `CREATE TABLE` możesz zobaczyć błąd `relation {nazwa_tabeli} already exists`. Oznacza to, że istnieje już tabela o danej nazwie. Musisz wtedy albo usunąć istniejącą tabelę o tej nazwie, ale zmienić nazwę nowej tabeli.

Teraz poznasz drugi sposób tworzenia tabel — za pomocą kwerend w SQL-u. Najpierw jednak wykonaj ćwiczenie, w którym utworzysz tabelę za pomocą SQL-a.

Ćwiczenie 1.07 — tworzenie tabeli w SQL-u

W tym ćwiczeniu utworzysz tabelę za pomocą instrukcji `CREATE TABLE`. Dział marketingu w firmie ZoomZoom chce utworzyć tabelę `countries`, aby przeanalizować dane dla różnych państw. Ta tabela ma zawierać cztery kolumny: całkowitoliczbową kolumnę z kluczem, kolumnę z unikatową nazwą, kolumnę z rokiem powstania i kolumnę ze stolicą.

Oto kroki niezbędne do wykonania tego zadania:

1. Otwórz wybranego klienta SQL-a i nawiąż połączenie z bazą `sqlda`.
2. Wykonaj poniższą kwerendę, aby usunąć tabelę `countries`, jeśli w bazie istnieje już tabela o tej nazwie:

   ```
   DROP TABLE IF EXISTS countries;
   ```

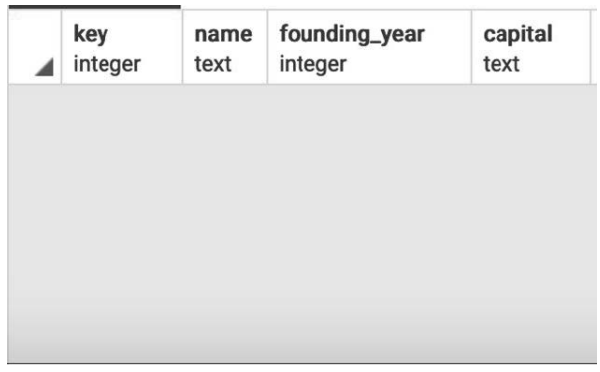
3. Uruchom następującą kwerendę, aby utworzyć tabelę `countries`:

   ```
   CREATE TABLE countries (
     key INT PRIMARY KEY,
     name text UNIQUE,
     founding_year INT,
     capital text
   );
   ```

   Powinieneś otrzymać pustą tabelę z rysunku 1.43.

> Kod źródłowy z tego punktu znajdziesz na stronie *https://packt.live/3cWFoSE*.

Kup książkę Poleć książkę



Rysunek 1.43. Pusta tabela countries z nazwami kolumn

W tym ćwiczeniu pokazaliśmy, jak utworzyć tabelę z różnymi ograniczeniami kolumn za pomocą instrukcji `CREATE TABLE`. W następnym punkcie utworzysz tabele przy użyciu kwerendy `SELECT`.

Tworzenie tabel za pomocą kwerendy SELECT

Wiesz już, jak utworzyć tabelę. Załóżmy jednak, że chcesz utworzyć tabelę na podstawie danych z istniejącej tabeli. Można to zrobić za pomocą innej wersji instrukcji `CREATE TABLE`:

```
CREATE TABLE {nazwa_tabeli} AS (
{kwerenda_select}
);
```

Tu `{kwerenda_select}` to dowolna kwerenda `SELECT`, jaką można uruchomić w bazie. Załóżmy, że chcesz utworzyć tabelę opartą na tabeli `products`, ale zawierającą tylko produkty z 2014 roku. Nazwij tę tabelę `products_2014`. Potrzebna kwerenda wygląda tak:

```
CREATE TABLE products_2014 AS (
SELECT
  *
FROM
  products
WHERE
  year=2014
);
```

W taki sposób można użyć dowolnej kwerendy, a nowa tabela odziedziczy wszystkie wartości z pierwotnej tabeli.

Kup książkę Poleć książkę

Aktualizowanie tabel

W przyszłości konieczne może być zmodyfikowanie tabeli przez dodanie kolumn, dodanie danych lub aktualizację istniejących wierszy. W tym podrozdziale zobaczysz, jak to zrobić.

Dodawanie i usuwanie kolumn

Aby dodać nowe kolumny do istniejącej tabeli, użyj instrukcji `ADD COLUMN`:

```
ALTER TABLE {nazwa_tabeli}
ADD COLUMN {nazwa_kolumny} {typ_danych};
```

Załóżmy, że chcesz dodać do tabeli `products` nową kolumnę, `weight`, która posłuży do zapisywania wagi produktów w kilogramach. Możesz to zrobić za pomocą następującej kwerendy:

```
ALTER TABLE products
ADD COLUMN weight INT;
```

Ta kwerenda tworzy w tabeli `products` nową kolumnę, `weight`, i przypisuje jej całkowitoliczbowy typ danych, aby można w niej było zapisywać tylko liczby.

Jeśli chcesz usunąć kolumnę z tabeli, możesz użyć instrukcji `DROP COLUMN`:

```
ALTER TABLE {nazwa_tabeli}
DROP COLUMN {nazwa_kolumny};
```

Tu `{nazwa_tabeli}` to nazwa tabeli, którą chcesz zmodyfikować, a `{nazwa_kolumny}` to nazwa usuwanej kolumny.

Jeśli chcesz usunąć wcześniej utworzoną kolumnę `weight`, możesz to zrobić za pomocą następującej kwerendy:

```
ALTER TABLE products
DROP COLUMN weight;
```

Dodawanie nowych danych

W SQL-u nowe dane możesz dodać do tabeli za pomocą kilku metod.

Jedna z nich polega na bezpośrednim wstawieniu wartości do tabeli za pomocą instrukcji `INSERT INTO…VALUES`. Oto struktura tej instrukcji:

```
INSERT INTO {table_name} (
  {kolumna_1], {kolumna_2}, ... {kolumna_ostatnia}
)
VALUES (
  {wartość_kolumny_1}, {wartość_kolumny_2},
    … {wartość_kolumny_ostatniej}
);
```


Kup książkę Poleć książkę

Tu {nazwa_tabeli} to nazwa tabeli, do której chcesz wstawić dane, {kolumna_1}, {kolumna_2}, …, {kolumna_ostatnia} to lista kolumn, do których mają trafić wartości, a {wartość_kolumny_1}, {wartość_kolumny_2}, …, {wartość_kolumny_ostatniej} to lista wartości wiersza, jakie kwerenda ma zapisać w tabeli. Jeśli w instrukcji INSERT jakieś istniejące kolumny tabeli zostaną pominięte, domyślnie umieszczone zostaną w nich wartości NULL.

Załóżmy, że chcesz wstawić do tabeli products nowy skuter. Możesz użyć następującej kwerendy:

```
INSERT INTO products (
  product_id, model, year,
  product_type, base_msrp,
  production_start_date, production_end_date
)
VALUES (
  13, 'Nimbus 5000', 2019,
  'scooter', 500.00,
  '2019-03-03', '2020-03-03'
);
```

Ta kwerenda odpowiednio modyfikuje tabelę products, co widać na rysunku 1.44.

product_id bigint	model text	year bigint	product_type text	base_msrp text	production_start_date timestamp without time zone	production_end_date timestamp without time zone
1	Lemon	2010	scooter	399.99	2010-03-03 00:00:00	2012-06-08 00:00:00
2	Lemon …	2011	scooter	799.99	2011-01-03 00:00:00	2011-03-30 00:00:00
3	Lemon	2013	scooter	499.99	2013-05-01 00:00:00	2018-12-28 00:00:00
4	Model …	2014	automobile	115,000.00	2014-06-23 00:00:00	2018-12-28 00:00:00
5	Blade	2014	scooter	699.99	2014-06-23 00:00:00	2015-01-27 00:00:00
6	Model S…	2015	automobile	65,500.00	2015-04-15 00:00:00	2018-10-01 00:00:00
7	Bat	2016	scooter	599.99	2016-10-10 00:00:00	[null]
8	Bat Limi…	2017	scooter	699.99	2017-02-15 00:00:00	[null]
9	Model E…	2017	automobile	35,000.00	2017-02-15 00:00:00	[null]
10	Model …	2017	automobile	85,750.00	2017-02-15 00:00:00	[null]
11	Model …	2019	automobile	95,000.00	2019-02-04 00:00:00	[null]
12	Lemon …	2019	scooter	349.99	2019-02-04 00:00:00	[null]
13	Nimbus…	2019	scooter	500.00	2019-03-03 00:00:00	2020-03-03 00:00:00

Rysunek 1.44. Tabela products po udanym wstawieniu kwerendy INSERT

Innym sposobem na wstawianie danych do tabeli jest użycie instrukcji INSERT razem z kwerendą SELECT. Oto składnia tego rozwiązania:

```
INSERT INTO {nazwa_tabeli} ({kolumna_1}, {kolumna_2}, ... {kolumna_ostatnia})
{kwerenda_select};
```

Tu {nazwa_tabeli} to nazwa tabeli, w której chcesz wstawić dane, {kolumna_1}, {kolumna_2}, …, {kolumna_ostatnia} to lista kolumn, w których chcesz umieścić wartości, a {kwerenda_select} to kwerenda o strukturze zgodnej ze strukturą wartości wstawianych do tabeli.

Kup książkę Poleć książkę

Przyjrzyj się opisanej wcześniej tabeli products_2014. Załóżmy, że zamiast tworzyć ją za pomocą kwerendy SELECT przygotowałeś pustą tabelę o tej samej strukturze co tabela products. Jeśli chcesz wstawić te same dane co wcześniej, możesz użyć następującej kwerendy:

```
INSERT INTO products_2014(
  product_id, model, year,
  product_type, base_msrp,
  production_start_date, production_end_date
)
SELECT
  *
FROM
  products
WHERE
  year=2014;
```

Ta kwerenda daje efekt widoczny na rysunku 1.45.

product_id bigint	model text	year bigint	product_type text	base_msrp text	production_start_date timestamp without time zone	production_end_date timestamp without time zone
4	Model ...	2014	automobile	115,000.00	2014-06-23 00:00:00	2018-12-28 00:00:00
5	Blade	2014	scooter	699.99	2014-06-23 00:00:00	2015-01-27 00:00:00

Rysunek 1.45. Tabela products_2014 po udanym wykonaniu kwerendy INSERT INTO

Teraz dowiesz się, jak zaktualizować zawartość wiersza.

Aktualizowanie istniejących wierszy

Czasem trzeba zaktualizować wartości danych już znajdujących się w tabeli. Możesz użyć do tego instrukcji UPDATE:

```
UPDATE {nazwa_tabeli}
SET {kolumna_1} = {wartość_kolumny_1},
    {kolumna_2} = {wartość_kolumny_2},
    …
    {kolumna_ostatnia} = {{wartość_kolumny_ostatniej}}
WHERE
    {warunek};
```

Tu {nazwa_tabeli} to nazwa tabeli z modyfikowanymi danymi, {kolumna_1}, {kolumna_2}, …, {kolumna_ostatnia} to lista kolumn, których wartości chcesz zmodyfikować, a {wartość_kolumny_1}, {wartość_kolumny_2}, …, {wartość_kolumny_ostatniej} to lista nowych wartości, które chcesz wstawić w tych kolumnach. Z kolei {warunek} to instrukcja warunkowa, taka, jakie stosuje się w kwerendach SQL.

Załóżmy, że przez resztę roku firma zamierza sprzedawać wszystkie skutery z roczników starszych niż 2018 po 299,99 dolara. Możesz zmodyfikować odpowiednie dane w tabeli products, używając następującej kwerendy:

```
UPDATE
  products
SET
```


Kup książkę Poleć książkę

```
  base_msrp = 299.99
WHERE
  product_type = 'scooter'
  AND year<2018;
```

Ta kwerenda zwraca dane wyjściowe widoczne na rysunku 1.46.

product_id bigint	model text	year bigint	product_type text	base_msrp text	production_start_date timestamp without time zone	production_end_date timestamp without time zone
1	Lemon	2010	scooter	299.99	2010-03-03 00:00:00	2012-06-08 00:00:00
2	Lemon ...	2011	scooter	299.99	2011-01-03 00:00:00	2011-03-30 00:00:00
3	Lemon	2013	scooter	299.99	2013-05-01 00:00:00	2018-12-28 00:00:00
4	Model ...	2014	automobile	115,000.00	2014-06-23 00:00:00	2018-12-28 00:00:00
5	Blade	2014	scooter	299.99	2014-06-23 00:00:00	2015-01-27 00:00:00
6	Model S...	2015	automobile	65,500.00	2015-04-15 00:00:00	2018-10-01 00:00:00
7	Bat	2016	scooter	299.99	2016-10-10 00:00:00	[null]
8	Bat Limi...	2017	scooter	299.99	2017-02-15 00:00:00	[null]
9	Model E...	2017	automobile	35,000.00	2017-02-15 00:00:00	[null]
10	Model ...	2017	automobile	85,750.00	2017-02-15 00:00:00	[null]
11	Model ...	2019	automobile	95,000.00	2019-02-04 00:00:00	[null]
12	Lemon ...	2019	scooter	349.99	2019-02-04 00:00:00	[null]
13	Nimbus...	2019	scooter	500.00	2019-03-03 00:00:00	2020-03-03 00:00:00

Rysunek 1.46. Udana aktualizacja tabeli products

W następnym ćwiczeniu przyjrzysz się dokładnie działaniu instrukcji `UPDATE` w bazie SQL-owej.

Ćwiczenie 1.08 — aktualizowanie tabeli w celu podniesienia ceny pojazdu

W tym ćwiczeniu za pomocą instrukcji `UPDATE` zaktualizujesz dane w tabeli. Z powodu wyższych kosztów rzadkich metali potrzebnych do produkcji samochodów elektrycznych cena nowego modelu Chi z 2019 roku musi zostać podniesiona o 10% względem aktualnej ceny 95 000 dolarów. Zaktualizuj tabelę `products`, aby podwyższyć cenę produktu.

Oto kroki niezbędne do wykonania tego zadania:

1. Otwórz wybranego klienta SQL-a, aby nawiązać połączenie z bazą `sqlda`.
2. Uruchom poniższą kwerendę, aby zaktualizować cenę samochodu `Model Chi` w tabeli `products`:

```
UPDATE
  products
SET
  base_msrp = base_msrp*1.10
WHERE
  model='Model Chi'
  AND year=2019;
```


Kup książkę Poleć książkę

3. Teraz napisz kwerendę `SELECT`, aby sprawdzić, czy cena samochodu `Model Chi` z 2019 roku została zaktualizowana:

```
SELECT
  *
FROM
  products
WHERE
  model='Model Chi'
  AND year=2019;
```

Na rysunku 1.47 pokazane są dane wyjściowe tego kodu.

product_id	model	year	product_type	base_msrp	production_start_date	production_end_date
11	Model Chi	2019	automobile	104500	2019-02-04 00:00:00	NULL

Rysunek 1.47. Zaktualizowana cena pojazdu Model Chi z 2019 roku.

W tych danych wyjściowych widać, że cena pojazdu `Model Chi` wynosi obecnie `104 500`, choć wcześniej była równa `95 000`.

> Kod źródłowy z tego punktu jest dostępny na stronie *https://packt.live/2XRJVl7*.

W tym ćwiczeniu pokazaliśmy, jak zaktualizować tabelę za pomocą instrukcji `UPDATE`. Teraz dowiesz się, jak usuwać tabele i dane.

Usuwanie danych i tabel

Często okazuje się, że dane w tabeli są nieprawidłowe, dlatego nie mogą być już używane. Wtedy trzeba usunąć dane z tabeli.

Usuwanie wartości z wiersza

Niekiedy trzeba usunąć wartość z wiersza. Najłatwiej użyć do tego opisanej już instrukcji `UPDATE` i przypisać do kolumny wartość `NULL`:

```
UPDATE {nazwa_tabeli}
SET {kolumna_1} = NULL,
    {kolumna_2} = NULL,
    …
    {kolumna_ostatnia} = NULL
WHERE
    {warunek};
```


Kup książkę Poleć książkę

W tej instrukcji {nazwa_tabeli} to nazwa tabeli, w której dane są modyfikowane, {kolumna_1}, {kolumna_2}, …, {kolumna_ostatnia} to lista kolumn, z których chcesz usunąć wartości, a {warunek} to instrukcja warunkowa, taka, jakie stosuje się w kwerendach SQL.

Załóżmy, że w pliku zapisany jest błędny adres e-mail klienta o identyfikatorze równym 3. Aby to zmienić, możesz użyć następującej kwerendy:

```
UPDATE
  customers
SET
  email = NULL
WHERE
  customer_id=3;
```

W następnym punkcie nauczysz się usuwać wiersze z tabeli.

Usuwanie wierszy z tabeli

Aby usunąć wiersz z tabeli, możesz użyć instrukcji DELETE. Wygląda ona tak:

```
DELETE FROM {nazwa_tabeli}
WHERE {warunek};
```

Załóżmy, że musisz usunąć szczegółowe informacje o kliencie z adresem e-mail bjordan2@geocities.com. Aby to zrobić, możesz się posłużyć następującą kwerendą:

```
DELETE FROM
  customers
WHERE
  email='bjordan2@geocities.com';
```

Jeśli chcesz usunąć wszystkie dane z tabeli customers bez kasowania jej samej, możesz zastosować taką kwerendę:

```
DELETE FROM customers;
```

Inny sposób na usunięcie w kwerendzie wszystkich danych bez kasowania tabeli polega na użyciu słowa kluczowego TRUNCATE:

```
TRUNCATE TABLE customers;
```

Usuwanie tabel

Aby usunąć wszystkie dane z tabeli i samą tabelę, możesz wywołać instrukcję DROP TABLE. Oto jej składnia:

```
DROP TABLE {nazwa_tabeli};
```

Tu {nazwa_tabeli} to nazwa tabeli, którą chcesz usunąć. Jeśli zamierzasz skasować wszystkie dane z tabeli customers wraz z samą tabelą, użyj takiej kwerendy:

```
DROP TABLE customers;
```

Teraz wykonaj ćwiczenie, aby usunąć tabelę za pomocą instrukcji DROP TABLE.

Kup książkę Poleć książkę

Ćwiczenie 1.09 — usuwanie niepotrzebnej tabeli

W tym ćwiczeniu nauczysz się usuwać tabelę za pomocą SQL-a. Dział marketingu zakończył analizowanie potencjalnej liczby klientów w każdym stanie i nie potrzebuje już tabeli `state_populations`. Aby zmniejszyć ilość miejsca zajmowanego przez bazę danych, usuń tę tabelę.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Otwórz wybranego klienta SQL-a i nawiąż połączenie z bazą `sqlda`.
2. Uruchom poniższą kwerendę, aby usunąć tabelę `state_populations`:

   ```
   DROP TABLE state_populations;
   ```

   Tabela `state_populations` powinna zostać usunięta z bazy danych.
3. Ponieważ tabela została skasowana, dotycząca jej kwerenda `SELECT` zwróci — zgodnie z oczekiwaniami — błąd:

   ```
   SELECT
     *
   FROM
     state_populations;
   ```

   Ten błąd jest pokazany na rysunku 1.48.

```
ERROR: BŁĄD:  relacja "state_populations" nie istnieje
LINE 4:   state_populations;
          ^
```

Rysunek 1.48. Błąd po usunięciu tabeli state_populations

> Kod źródłowy z tego punktu jest dostępny pod adresem *https://packt.live/2XWLVZA*.

W tym ćwiczeniu pokazaliśmy, jak usunąć tabelę za pomocą instrukcji `DROP TABLE`. W następnym zadaniu utworzysz i zmodyfikujesz tabele za pomocą SQL-a.

Zadanie 1.04 — tworzenie i modyfikowanie tabel na potrzeby działań marketingowych

W tym zadaniu przetestujesz swoje umiejętności z zakresie tworzenia i modyfikowania tabel za pomocą SQL-a. Wykonałeś dobrą robotę w pobieraniu danych dla działu marketingu. Jednak menedżer ds. marketingu, któremu pomogłeś, zdał sobie sprawę, że o czymś zapomniał. Okazuje się, że oprócz pobrania danych chce też utworzyć nową tabelę w analitycznej bazie danych firmy. Ponadto musi wprowadzić zmiany w danych z tabeli `customers`. Twoje zadanie polega na tym, by pomóc menedżerowi w tworzeniu tabeli:

Kup książkę Poleć książkę

1. Utwórz nową tabelę `customers_nyc`, która pobiera z tabeli `customers` wszystkie wiersze dotyczące klientów mieszkających w mieście Nowy Jork w stanie o tej samej nazwie.
2. Usuń z nowej tabeli wszystkich klientów, którzy mają adres z kodem pocztowym `10014`. Z powodu lokalnych przepisów nie można kierować do nich akcji marketingowej.
3. Dodaj nową kolumnę tekstową `event`.
4. Ustaw wartość w kolumnie `event` na `Impreza w ramach podziękowań`. Na rysunku 1.49 pokazane są oczekiwane dane wyjściowe.

	customer_id bigint	title text	first_name text	last_name text	suffix text	email text	gender text	ip_address text	phone text	street_address text	city text	state text	postal_co text	latitude double	longitude double	date_added timestamp	event text
1	52	[null]	Giusto	Backe	[null]	gback...	M	26.56.68.189	212-9...	6 Onsgard Terrace	New York City	NY	10131	40.7808	-73.9772	2010-07-06 ...	Impreza w ramach podziękowań
2	162	[null]	Artair	Betchley	[null]	abetc...	M	108.147.128.250	[null]	7 Boyd Road	New York City	NY	10090	40.7808	-73.9772	2014-06-25 ...	Impreza w ramach podziękowań
3	406	[null]	Rozina	Jeal	[null]	rjealb...	F	50.235.32.29	917-6...	64653 Homewood T...	New York City	NY	10105	40.7628	-73.9785	2010-09-15 ...	Impreza w ramach podziękowań
4	456	Rev	Cybil	Noke	[null]	cnoke...	F	5.31.139.106	212-3...	88 Sycamore Parkw...	New York City	NY	10280	40.7808	-73.9772	2017-01-21 ...	Impreza w ramach podziękowań
5	472	[null]	Rawley	Yegorov	[null]	ryego...	M	183.199.243.74	212-5...	872 Old Shore Park...	New York City	NY	10034	40.8662	-73.9221	2014-11-24 ...	Impreza w ramach podziękowań
6	496	[null]	Layton	Spolton	[null]	lspolt...	M	108.112.8.165	646-9...	7 Old Gate Drive	New York City	NY	10024	40.7864	-73.9764	2010-12-20 ...	Impreza w ramach podziękowań
7	1028	[null]	Issy	Andrieux	[null]	iandri...	F	199.50.5.37	212-2...	33337 Dahle Way	New York City	NY	10115	40.8111	-73.9642	2017-11-27 ...	Impreza w ramach podziękowań
8	1037	[null]	Magdalene	Veryard	[null]	mvery...	F	93.201.129.213	[null]	41028 Katie Junction	New York City	NY	10039	40.8265	-73.9383	2014-03-04 ...	Impreza w ramach podziękowań
9	1063	[null]	Juliet	Beadles	[null]	jbeadl...	F	47.96.88.226	212-6...	34984 Goodland Poi...	New York City	NY	10120	40.7506	-73.9894	2014-08-17 ...	Impreza w ramach podziękowań
10	1211	[null]	Gwyneth	McCobb	[null]	gmcc...	F	38.182.151.212	[null]	4 Jana Park	New York City	NY	10160	40.7808	-73.9772	2014-01-08 ...	Impreza w ramach podziękowań
11	1262	[null]	Conrado	Escoffier	[null]	cesco...	M	23.120.12.44	646-5...	2 Atwood Court	New York City	NY	10060	40.7808	-73.9772	2015-02-17 ...	Impreza w ramach podziękowań
12	1333	[null]	Franny	Tipping	[null]	ftippi...	F	125.115.139.118	212-5...	93444 Drewry Trail	New York City	NY	10292	40.7808	-73.9772	2014-09-17 ...	Impreza w ramach podziękowań
13	1403	[null]	Terza	Dericut	[null]	tderic...	F	243.18.169.116	646-7...	3037 Linden Center	New York City	NY	10060	40.7808	-73.9772	2015-07-10 ...	Impreza w ramach podziękowań
14	1587	[null]	Gregoire	Marciek	[null]	gmar...	M	75.31.7.169	[null]	9 Lunder Place	New York City	NY	10039	40.8265	-73.9383	2011-01-14 ...	Impreza w ramach podziękowań
15	1675	[null]	Ram	Acheson	[null]	rache...	M	34.129.78.67	718-4...	5 Rutledge Point	New York City	NY	10019	40.7651	-73.9858	2016-06-15 ...	Impreza w ramach podziękowań
16	1860	[null]	Fernanda	Haney	[null]	fhane...	F	148.198.91.104	[null]	17 Hanover Point	New York City	NY	10150	40.7808	-73.9772	2012-03-27 ...	Impreza w ramach podziękowań

Rysunek 1.49. Tabela customers_nyc z kolumną event o wartości Impreza w ramach podziękowań

5. Informujesz menedżera, że wykonałeś opisane kroki. Menedżer powiadamia o tym zespół ds. działań marketingowych, który wykorzystuje dane do przeprowadzenia kampanii. Menedżer dziękuje Ci, po czym prosi, abyś usunął tabelę `customers_nyc`.

> Rozwiązanie tego zadania znajdziesz w „Dodatku”.

W tym zadaniu wykorzystałeś różne operacje CRUD, aby zmodyfikować tabelę zgodnie z prośbami menedżera ds. marketingu. Po tym wstępie pora wrócić do powiązań SQL-a z analizą danych.

SQL i analityka

Możliwe, że w tym rozdziale dostrzegłeś związki między tabelami z SQL-a a zbiorami danych. Powinno być już zrozumiałe, że tabele z SQL-a należy traktować jak zbiory danych, wiersze to odpowiednik jednostek obserwacji, a kolumny przechowują wartości cech. Jeśli spojrzysz na tabele z SQL-a w ten sposób, zobaczysz, że SQL to naturalne narzędzie do przechowywania zbiorów danych w komputerze.

Kup książkę Poleć książkę

Jednak SQL nie tylko zapewnia wygodny sposób przechowywania zbiorów danych. Nowe implementacje SQL-a udostępniają też narzędzia do przetwarzania i analizowania danych z użyciem różnych funkcji. Za pomocą SQL-a możesz oczyszczać dane, przetwarzać je na bardziej przydatne formaty, a także analizować z wykorzystaniem statystyk, by wykrywać ciekawe wzorce. Reszta książki pozwoli Ci zrozumieć, jak korzystać z SQL-a do wykonywania takich operacji w produktywny i wydajny sposób.

Podsumowanie

Analiza danych to rozbudowana dziedzina pozwalająca zrozumieć świat. Ostatecznym celem analiz jest przekształcanie danych w informacje i wiedzę. Aby osiągnąć te cele, można posłużyć się statystyką (przede wszystkim statystyką opisową i testami istotności statystycznej), by lepiej zrozumieć dane.

Gałąź statystyki opisowej zwana analizą jednoczynnikową pomaga zrozumieć jedną zmienną z danych. Analiza jednoczynnikowa pozwala znajdować wartości odstające, badać rozkład danych za pomocą rozkładu częstości występowania i kwantyli, sprawdzać tendencję centralną na podstawie obliczeń średniej, mediany i wartości modalnej, a także analizować dyspersję danych z użyciem rozstępu, odchylenia standardowego i rozstępu ćwiartkowego.

Do zrozumienia relacji między danymi można też wykorzystać analizę dwuczynnikową. Za pomocą wykresów punktowych można ocenić trendy, zmiany w trendach, zjawiska cykliczne i anomalie dotyczące dwóch zmiennych. Można także za pomocą współczynnika korelacji Pearsona zmierzyć siłę trendu liniowego dla dwóch zmiennych. Współczynnik ten należy jednak dokładnie sprawdzić pod kątem wartości odstających i liczby punktów danych użytych w obliczeniach. Ponadto wysoka korelacja między dwiema zmiennymi nie oznacza, że jedna zmienna przyczynowo wpływa na drugą.

Także testy istotności statystycznej mogą zapewniać ważne informacje na temat danych. Te testy pozwalają ocenić, jak prawdopodobne jest przypadkowe uzyskanie określonych wyników. Pomagają też zrozumieć, czy zmiany i różnice między grupami są istotne statystycznie.

Analizę danych można wzbogacić dzięki możliwościom, jakie dają relacyjne bazy danych. Bazy relacyjne to dojrzała i wszechobecna technologia przechowywania i pobierania danych. Takie bazy przechowują dane w formie relacji (nazywanych tabelami) i zapewniają doskonałe połączenie szybkości działania, wydajności i łatwości użytkowania. SQL to język używany do dostępu do relacyjnych baz danych. Jest językiem deklaratywnym, dlatego umożliwia użytkownikom skupienie się na tym, co chcą uzyskać, a nie na procesie tworzenia wyniku. SQL udostępnia wiele typów danych, w tym typy liczbowe i tekstowe, a nawet struktury danych.

W procesie pobierania danych SQL umożliwia użytkownikowi wskazanie pól, z których dane należy wczytać, a także sposobu filtrowania danych. Dane można też sortować. Ponadto SQL umożliwia pobranie dowolnie dużej lub małej ilości danych. Również operacje tworzenia, wczytywania, aktualizowania i usuwania danych są dość proste i można je precyzyjnie wykonywać.

Po objaśnieniu podstaw analizy danych i SQL-a w następnym rozdziale przejdziemy do omówienia tego, jak używać SQL-a do wykonywania pierwszego kroku w analizie danych — oczyszczania i przekształcania danych.

Kup książkę Poleć książkę

Kup książkę Poleć książkę

Skorowidz

Kup książkę Poleć książkę

E

F

Kup książkę Poleć książkę

G

H

I

J

K

Kup książkę Poleć książkę

Ł

M

N

O

P

Kup książkę Poleć książkę

R

S

Ś

Kup książkę Poleć książkę

T

U

W

Kup książkę Poleć książkę

Z

Kup książkę Poleć książkę

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

SQL: ZNAKOMITE NARZĘDZIE W PROFESJONALNEJ ANALIZIE DANYCH!

Obecnie mamy dostęp do terabajtów danych. To nieprzebrane źródło cennych informacji, które mogą decydować o upadku albo o rozkwicie firmy. Aby jednak wydobyć z danych potrzebną wiedzę, trzeba się wykazać kompetencjami. Są to cenne umiejętności — profesjonalny analityk danych może przebierać w atrakcyjnych ofertach pracy. Spośród różnych technik analizy danych warto się przyjrzeć zastosowaniu zapytań SQL. SQL to język tworzony i rozwijany dla potrzeb pracy z bazami danych, jest więc szczególnie przydatnym narzędziem w przyborniku analityka danych.

Ta książka jest znakomitym przewodnikiem dla początkującego analityka danych. Dzięki niej dowiesz się, jak skutecznie przesiewać i uzyskiwać informacje z surowych danych. Nauczysz się formułować hipotezy i generować opisowe statystyki, a także pisać złożone zapytania SQL, które pozwalają na zagregowanie danych z bazy SQL z danymi pochodzącymi z innych źródeł. Zobaczysz, jak pracować z danymi w różnych formatach, nauczysz się analizy geoprzestrzennej i analizy tekstu. Poznasz też tajniki pozyskiwania informacji z wykorzystaniem takich metod jak profilowanie i automatyzacja.

W KSIĄŻCE:

- przygotowanie danych za pomocą zapytań SQL
- funkcje agregujące i funkcje okna w SQL
- bazy danych i Excel oraz kod w R i w Pythonie
- praca ze złożonymi typami danych
- optymalizacja zapytań SQL
- metodyczne rozwiązywanie problemów

Matt Goldwasser od lat pracuje jako analityk danych w prestiżowych firmach. Lubi rozwiązywać problemy z uczeniem maszynowym i poznawać nowe technologie.
Upom Malik jest naukowcem, zajmuje się analizą danych i stosowaniem SQL do rozwiązywania problemów z branży finansów i energetyki.
Benjamin Johnston zajmuje się zaawansowaną analizą danych w branży medycznej. Interesuje się uczeniem maszynowym, przetwarzaniem obrazów i sieciami neuronowymi.

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
HELIONSZKOLENIA.PL

KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-283-8474-3
9 788328 384743
INFORMATYKA W NAJLEPSZYM WYDANIU
Cena: 89,00 zł
Packt>