SQL
DLA ANALITYKÓW DANYCH
WYDANIE III

Opanuj możliwości SQL-a,
aby wydobywać informacje
z danych

Jun Shan | Matt Goldwasser | Upom Malik | Benjamin Johnston

Packt›

Helion

Tytuł oryginału: SQL for Data Analytics: Harness the power of SQL to extract insights from data, 3rd Edition

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-289-0173-5

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/sqlan3
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

Wprowadzenie **11**

ROZDZIAŁ 1
Poznawanie i opisywanie danych **37**

Wprowadzenie 37
Analityka danych i statystyka 38
Zadanie 1.01 — klasyfikowanie nowego zbioru danych 39
Rodzaje statystyki 40
Metody z obszaru statystyki opisowej 41
Analiza jednoczynnikowa 41
Ćwiczenie 1.01 — tworzenie histogramu 42
Ćwiczenie 1.02 — obliczanie kwartyli dla sprzedaży dodatków 50
Ćwiczenie 1.03 — obliczanie miar tendencji centralnej dla sprzedaży dodatków 54
Ćwiczenie 1.04 — obliczanie dyspersji dla sprzedaży dodatków 57
Analiza dwuczynnikowa 58
Ćwiczenie 1.05 — obliczanie współczynnika korelacji Pearsona dla dwóch zmiennych 64
Interpretowanie i analizowanie współczynnika korelacji 66
Zadanie 1.02 — eksplorowanie danych sprzedażowych z salonu samochodowego 70
Praca z niepełnymi danymi 70
Testy istotności statystycznej 71
Często używane testy istotności statystycznej 72
SQL i analityka 73
Podsumowanie 73

ROZDZIAŁ 2
Wprowadzenie do SQL-a dla analityków **75**

Wprowadzenie 75
Świat danych 76
Rodzaje danych 77
Relacyjne bazy danych i SQL 78
Wady i zalety baz SQL-owych 79

System zarządzania relacyjnymi bazami danych PostgreSQL 81
Ćwiczenie 2.01 — uruchamianie pierwszej kwerendy SELECT 83
Instrukcja SELECT 86
Klauzula WHERE 90
Klauzule AND i OR 91
Klauzule IN i NOT IN 93
Klauzula ORDER BY 94
Klauzula LIMIT 97
Klauzule IS NULL i IS NOT NULL 98
Ćwiczenie 2.02 — kwerenda SELECT z podstawowymi słowami kluczowymi dotycząca tabeli salespeople 99
Zadanie 2.01 — kwerenda SELECT z podstawowymi słowami kluczowymi dotycząca tabeli customers 102
Tworzenie tabel 103
Tworzenie pustych tabel 103
Podstawowe typy danych w SQL-u 104
Typy liczbowe 104
Typy znakowe 104
Typ logiczny 105
Daty i godziny 106
Struktury danych — format JSON i tablice 106
Ograniczenia kolumn 107
Prosta instrukcja CREATE 107
Ćwiczenie 2.03 — tworzenie tabeli w SQL-u 108
Tworzenie tabel za pomocą kwerendy SELECT 109
Aktualizowanie tabel 111
Dodawanie i usuwanie kolumn 111
Dodawanie nowych danych 112
Aktualizowanie istniejących wierszy 114
Ćwiczenie 2.04 — aktualizowanie tabeli w celu podniesienia ceny pojazdu 115
Usuwanie danych i tabel 116
Usuwanie wartości z wiersza 116
Usuwanie wierszy z tabeli 117
Usuwanie tabel 117
Ćwiczenie 2.05 — usuwanie niepotrzebnej tabeli pomocniczej 119
Zadanie 2.02 — tworzenie i modyfikowanie tabel na potrzeby działań marketingowych 120
SQL i analityka 121
Podsumowanie 121

ROZDZIAŁ 3
Przygotowywanie danych za pomocą SQL-a 123
Wprowadzenie 123
Łączenie danych 124
Łączenie tabel za pomocą słowa kluczowego JOIN 124
Rodzaje złączeń 127
Ćwiczenie 3.01 — używanie złączeń do analizy sprzedaży w salonach 137
Podkwerendy 139
Sumy 140
Ćwiczenie 3.02 — generowanie listy gości na przyjęcie dla klientów VIP za pomocą klauzuli UNION 142
Wyrażenia WITH 144
Oczyszczanie danych 145
Funkcja CASE WHEN 145
Ćwiczenie 3.03 — używanie funkcji CASE WHEN do pobierania list klientów z danego regionu 147
Funkcja COALESCE 148
Funkcja NULLIF 149
Funkcje LEAST i GREATEST 151
Funkcja CASTING 152
Przekształcanie danych 153
Funkcje DISTINCT i DISTINCT ON 153
Zadanie 3.01 — używanie SQL-a do tworzenia modelu wspomagającego sprzedaż 156
Podsumowanie 157

ROZDZIAŁ 4
Analiza danych z wykorzystaniem funkcji agregujących 159
Wprowadzenie 159
Funkcje agregujące 160
Ćwiczenie 4.01 — używanie funkcji agregujących do analizowania danych 165
Funkcje agregujące z klauzulą GROUP BY 166
Klauzula GROUP BY 167
Klauzula GROUP BY dla kilku kolumn 173
Ćwiczenie 4.02 — obliczanie cen dla typów produktów za pomocą klauzuli GROUP BY 174
Klauzula GROUPING SETS 175
Funkcje agregujące dla zbiorów uporządkowanych 177
Funkcje agregujące z klauzulą HAVING 178
Ćwiczenie 4.03 — obliczanie wyników i wyświetlanie danych z użyciem klauzuli HAVING 180

Stosowanie funkcji agregujących do oczyszczania danych i sprawdzania ich jakości ... 181
Znajdowanie brakujących wartości za pomocą klauzuli GROUP BY ... 181
Sprawdzanie unikatowości danych za pomocą funkcji agregujących ... 184
Zadanie 4.01 — analizowanie danych sprzedażowych z użyciem funkcji agregujących ... 185
Podsumowanie ... 186

ROZDZIAŁ 5
Analizowanie danych za pomocą funkcji okna ... 188
Wprowadzenie ... 188
Funkcje okna ... 189
Podstawy funkcji okna ... 191
Ćwiczenie 5.01 — analizowanie zmian współczynnika podawania danych przez klientów w czasie ... 197
Słowo kluczowe WINDOW ... 200
Obliczanie statystyk z użyciem funkcji okna ... 202
Ćwiczenie 5.02 — określanie pozycji na podstawie daty rejestracji ... 203
Ramka okna ... 204
Ćwiczenie 5.03 — motywowanie pracowników lunchem ... 207
Zadanie 5.01 — analizowanie sprzedaży z wykorzystaniem ramek okna i funkcji okna ... 209
Podsumowanie ... 211

ROZDZIAŁ 6
Importowanie i eksportowanie danych ... 212
Wprowadzenie ... 212
Polecenie COPY ... 213
Uruchamianie polecenia psql ... 214
Instrukcja COPY ... 215
Instrukcja \COPY w narzędziu psql ... 218
Tworzenie tymczasowych widoków ... 219
Konfigurowanie poleceń COPY i \COPY ... 221
Użycie poleceń COPY i \COPY do masowego wczytywania danych do bazy ... 222
Ćwiczenie 6.01 — eksportowanie danych do pliku w celu dalszego przetwarzania ich w Excelu ... 223
Zastosowanie języka Python do bazy danych ... 228
Wprowadzenie do języka Python ... 228
Ułatwianie dostępu do baz PostgreSQL w Pythonie za pomocą narzędzi SQLAlchemy i pandas ... 233
Czym jest SQLAlchemy? ... 233
Stosowanie Pythona z wykorzystaniem pakietów SQLAlchemy i pandas ... 235

Pobieranie danych z bazy i ich zapisywanie w bazie za pomocą pakietu pandas ... 237
Zapisywanie danych w bazie za pomocą Pythona ... 238
Ćwiczenie 6.02 — wczytywanie, wizualizowanie i zapisywanie danych za pomocą Pythona ... 238
Zwiększanie szybkości zapisu w Pythonie za pomocą polecenia COPY ... 243
Odczyt i zapis plików CSV w Pythonie ... 245
Najlepsze praktyki z obszaru importowania i eksportowania danych ... 247
Pomijanie podawania hasła ... 247
Zadanie 6.01 — używanie zewnętrznego zbioru danych do wykrywania trendów sprzedażowych ... 248
Podsumowanie ... 250

ROZDZIAŁ 7
Analizy z wykorzystaniem złożonych typów danych ... 251
Wprowadzenie ... 251
Wykorzystywanie typów danych z datami i czasem do analiz ... 252
Wprowadzenie do typu date ... 252
Przekształcanie typów danych ... 255
Przedziały ... 257
Ćwiczenie 7.01 — analiza danych z szeregów czasowych ... 259
Przeprowadzanie analiz geoprzestrzennych w PostgreSQL ... 261
Długość i szerokość geograficzna ... 262
Reprezentowanie długości i szerokości geograficznej w PostgreSQL ... 262
Ćwiczenie 7.02 — analizy geoprzestrzenne ... 265
Stosowanie tablicowych typów danych w PostgreSQL ... 268
Wprowadzenie do tablic ... 268
Ćwiczenie 7.03 — analizowanie sekwencji z użyciem tablic ... 271
Stosowanie formatu JSON w PostgreSQL ... 274
JSONB — wstępnie przetworzone dane w formacie JSON ... 277
Dostęp do danych z pól w formacie JSON lub JSONB ... 277
Stosowanie języka JSONPath do pól w formacie JSONB ... 280
Tworzenie i modyfikowanie danych w polu w formacie JSONB ... 282
Ćwiczenie 7.04 — przeszukiwanie obiektów JSONB ... 283
Analiza tekstu za pomocą PostgreSQL ... 286
Tokenizacja tekstu ... 286
Ćwiczenie 7.05 — analizowanie tekstu ... 288
Wyszukiwanie tekstu ... 293
Optymalizowanie wyszukiwania tekstu w PostgreSQL ... 296
Zadanie 7.01 — wyszukiwanie i analiza transakcji sprzedaży ... 298
Podsumowanie ... 300

ROZDZIAŁ 8
Wydajny SQL 301
Wprowadzenie 301
Znaczenie wysoce wydajnego kodu w SQL-u 302
Metody skanowania baz danych 303
Plany wykonywania kwerend 304
Ćwiczenie 8.01 — interpretowanie działania planera kwerend 305
Zadanie 8.01 — plany wykonywania kwerendy 309
Skanowanie indeksu 310
Indeks w postaci B-drzewa 311
Ćwiczenie 8.02 — tworzenie indeksu na potrzeby skanowania 312
Zadanie 8.02 — stosowanie skanowania indeksu 317
Indeks z haszowaniem 318
Ćwiczenie 8.03 — tworzenie kilku indeksów z haszowaniem, aby zbadać ich wydajność 320
Zadanie 8.03 — stosowanie indeksów z haszowaniem 323
Skuteczne korzystanie z indeksów 324
Kończenie pracy kwerend 325
Ćwiczenie 8.04 — anulowanie długo działającej kwerendy 326
Funkcje i wyzwalacze 328
Definicje funkcji 328
Ćwiczenie 8.05 — tworzenie funkcji, które nie przyjmują argumentów 329
Zadanie 8.04 — definiowanie funkcji zwracającej maksymalną wartość sprzedaży 332
Ćwiczenie 8.06 — tworzenie funkcji przyjmujących argumenty 333
Zadanie 8.05 — tworzenie funkcji przyjmujących argumenty 335
Wyzwalacze 336
Ćwiczenie 8.07 — tworzenie wyzwalaczy do aktualizowania pól 339
Zadanie 8.06 — tworzenie wyzwalacza do śledzenia średniej liczby kupionych sztuk 344
Podsumowanie 345

ROZDZIAŁ 9
Odkrywanie prawdy za pomocą SQL-a — studium przypadku 346
Wprowadzenie 346
Studium przypadku 347
Metoda naukowa 347
Ćwiczenie 9.01 — wstępne zbieranie danych za pomocą technik SQL-a 348
Ćwiczenie 9.02 — pobieranie informacji sprzedażowych 351
Zadanie 9.01 — ilościowa ocena spadku sprzedaży 356

Ćwiczenie 9.03 — analiza czasu rozpoczęcia sprzedaży 358
Zadanie 9.02 — analiza hipotezy dotyczącej różnicy w cenie sprzedaży 367
Ćwiczenie 9.04 — analiza zależności wzrostu sprzedaży od współczynnika otwarć e-maili 368
Ćwiczenie 9.05 — analiza skuteczności e-mailowej kampanii marketingowej 377
Wnioski 382
Badania terenowe 382
Podsumowanie 383

Dodatek 385

Skorowidz 429

Rozdział 1
Poznawanie i opisywanie danych

Omówienie rozdziału

Gdy zakończysz lekturę tego rozdziału, będziesz umieć opisywać dane i statystyki 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.

Wprowadzenie

Zbieranie i analizowanie danych to praktyka stosowana już od zarania cywilizacji. Zapisy ze starożytnego egipskiego papirusu wskazują, że faraonowie zbierali informacje ze spisów ludności wsi, co prawdopodobnie miało służyć określeniu liczby żołnierzy, których można powołać w razie wojny. Jednak dopiero po pojawieniu się nowoczesnych komputerów sztuka analizy danych stała się ważnym zjawiskiem, które każdego dnia zmienia życie ludzi.

Ta książka, jak wskazuje na to jej tytuł, uczy stosowania języka **SQL** (ang. *Structured Query Language*) do analizy danych. SQL to główne narzędzie używane w tej książce. Jednak zanim do niego przejdziemy, w tym rozdziale przedstawiamy wprowadzenie do analizy danych. Poznasz tu podstawowe zagadnienia, takie jak definicje i typy statystyk czy różne metody statystyczne. Prezentujemy podstawy, na których oparte są zagadnienia omawiane w dalszych rozdziałach, wyjaśniamy przeznaczenie operacji SQL-owych, a także przedstawiamy dziedzinę analityki, w ramach której są wykonywane te operacje. Ten rozdział rozpoczniesz od zapoznania się z danymi i statystykami.

Analityka danych i statystyka

Surowe dane są jedynie zestawem wartości, które można pobrać z ich źródła. Stają się przydatne dopiero wtedy, gdy są przetwarzane w celu znalezienia w nich różnych wzorców. Takie wzorce w danych to informacje. Pomagają one interpretować dane, przedstawiać prognozy i identyfikować nieoczekiwane przyszłe zmiany. Te informacje są następnie przekształcane w wiedzę.

Wiedza jest oparta na dużych uporządkowanych kolekcjach trwałych i rozbudowanych informacji oraz doświadczeń, które można wykorzystywać do opisywania i prognozowania zjawisk z rzeczywistego świata. 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 nich jest statystyka, w której do zbiorów danych używane są techniki matematyczne.

Statystyka to nauka związana ze zbieraniem i analizowaniem dużych ilości danych w celu identyfikowania cech tych danych i ich podzbiorów. Możesz na przykład zbadać historię danych medycznych z określonego kraju, aby zidentyfikować najczęstsze przyczyny zgonów wynikających z chorób. Możesz też dokładniej przyjrzeć się niektórym podgrupom (na przykład osobom z różnych obszarów geograficznych), aby ustalić, czy wśród ludzi zamieszkujących poszczególne regiony widoczne są specyficzne wzorce.

Statystyka jest stosowana do zbiorów danych. Dane w zbiorach mają różne cechy i wymagają różnych metod przetwarzania. Niektóre rodzaje danych, na przykład nazwy i etykiety, są **jakościowe**, co oznacza, że zawierają informacje opisowe. Inne dane, na przykład liczby i ilości, mają charakter **ilościowy**, co oznacza, że umożliwiają wykonywanie operacji liczbowych, takich jak dodawanie lub mnożenie. Na przykład zbiór danych z rysunku 1.1 to kolekcja informacji biomedycznych dotyczących grupy pacjentów.

W tym zbiorze danych jednostką obserwacji jest pojedynczy pacjent, ponieważ każdy wiersz reprezentuje jedną obserwację odpowiadającą unikatowemu pacjentowi. Występuje tu dziesięć punktów danych po pięć zmiennych w każdym. Trzy kolumny, `Data Urodzenia`, `Wzrost (cm)` i `Liczba Wizyt Lekarza w 2018`, są ilościowe, ponieważ do ich reprezentowania służą liczby. Dwie kolumny, `Kolor Oczu` i `Kraj Urodzenia`, są jakościowe.

Aby pomóc Ci zaznajomić się z podstawowymi zagadnieniami z obszaru zbiorów danych i statystyki, zamieszczamy ćwiczenie.

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 1.1. Dane dotyczące opieki zdrowotnej

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 1.2.

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 1.2. Dane na temat sprzedaży używanych samochodów

Oto kroki, jakie należy wykonać:

1. Ustalić jednostkę obserwacji.
2. Ocenić kolumny pod kątem tego, czy zawierają dane ilościowe, czy jakościowe. Użyj do tego definicji przedstawionej w tekście bezpośrednio przed zadaniem. Jeśli możesz wykonać na danych operacje arytmetyczne, dane są ilościowe, w przeciwnym razie są jakościowe.
3. Jeśli kolumna zawiera dane tekstowe, które są stałe i obejmują ograniczoną liczbę wartości, użyj wartości liczbowych do reprezentowania łańcuchów znaków. Jest to często stosowana technika, która przyspiesza przetwarzanie danych przez komputery. Na przykład aby przetwarzać dane w kolumnie z dniami tygodnia, możesz użyć 0 do reprezentowania niedzieli, 1 do reprezentowania poniedziałku itd. Spróbuj zastosować tę technikę i przekształcić kolumnę `Marka` na kolumnę z danymi ilościowymi.

W tym zadaniu nauczyłeś się klasyfikować dane. W następnym podrozdziale poznasz różne rodzaje statystyki.

Uwaga

Rozwiązanie tego zadania znajdziesz w „Dodatku”.

Rodzaje statystyki

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

Statystyka opisowa służy do opisywania kolekcji danych. Na przykład średni wiek mieszkańców jakiegoś państwa to wskaźnik statystyki opisowej reprezentujący cechę tych osób. Statystyki opisowe dotyczące jednej zmiennej ze zbioru danych to analizy **jednoczynnikowe**, a statystyki opisowe dotyczące jednocześnie dwóch lub więcej zmiennych to statystyki **wieloczynnikowe**. Statystyki dotyczące dwóch zmiennych są nazywane **analizami dwuczynnikowymi**. Średni wiek mieszkańców to wynik analizy jednoczynnikowej, a analizy badające zależności między produktem krajowym brutto na mieszkańca, wydatkami na służbę zdrowia na mieszkańca i wiekiem to analizy wieloczynnikowe.

Z kolei we **wnioskowaniu statystycznym** zbiór danych jest traktowany jako próbka, czyli niewielka część pomiarów z większej grupy nazywanej populacją. Na przykład ankieta obejmująca 10 000 w kraju zamieszkiwanym przez 100 milionów ludzi daje próbkę całej populacji. Zamiast sprawdzać wiek każdej osoby w państwie, można przepytać 10 000 ludzi i użyć ich średniego wieku jako średniej dla całego kraju.

Uwaga

W tej książce skupiamy się przede wszystkim na statystykach opisowych.

Metody z obszaru statystyki opisowej

W tym podrozdziale przyjrzysz się podstawowym technikom matematycznym analizy jedno- i dwuczynnikowej, które możesz wykorzystać, aby lepiej zrozumieć i opisać zbiór danych. Poznasz tu następujące metody w przedstawionej kolejności:

Techniki analizy jednoczynnikowej

- rozkład częstości,
- kwantyle,
- miary tendencji centralnej,
- dyspersja.

Techniki analizy dwuczynnikowej

- wykresy punktowe,
- liniowe analizy trendu i współczynnik korelacji Pearsona,
- interpretowanie i analizowanie współczynnika korelacji,
- szeregi czasowe.

Analiza jednoczynnikowa

W poprzednim podrozdziale wspomnieliśmy, że 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** (ponieważ używasz liczb bezwzględnych do przedstawienia częstości występowania określonego wzorca).

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. Osoba o wzroście 172 cm trafia do kubełka z przedziałem 170 – 174,99, a osoba o wzroście 181 cm zostanie przypisana do kubełka z przedziałem 180 – 184,99. Następnie można utworzyć rozkład zgodnie z liczbą elementów w kubełkach określonych na podstawie wzrostu. Taki rozkład jest oparty na bezwzględnej liczbie elementów w każdym kubełku, jest to więc rozkład bezwzględnej częstości występowania. Następnie można podzielić każdy wiersz tabeli przez łączną liczbę punktów danych i otrzymać rozkład względnej częstości występowania.

Inną przydatną techniką jest tworzenie wykresów rozkładów, czyli wizualizacji danych. Wizualizacje przedstawiają zależności między punktami danych w wizualnej formie, co ułatwia dostrzeżenie wzorców. W ćwiczeniu 1.01 utworzysz histogram, który jest graficzną reprezentacją rozkładu ciągłego z użyciem kubełków.

Ćwiczenie 1.01 — tworzenie histogramu

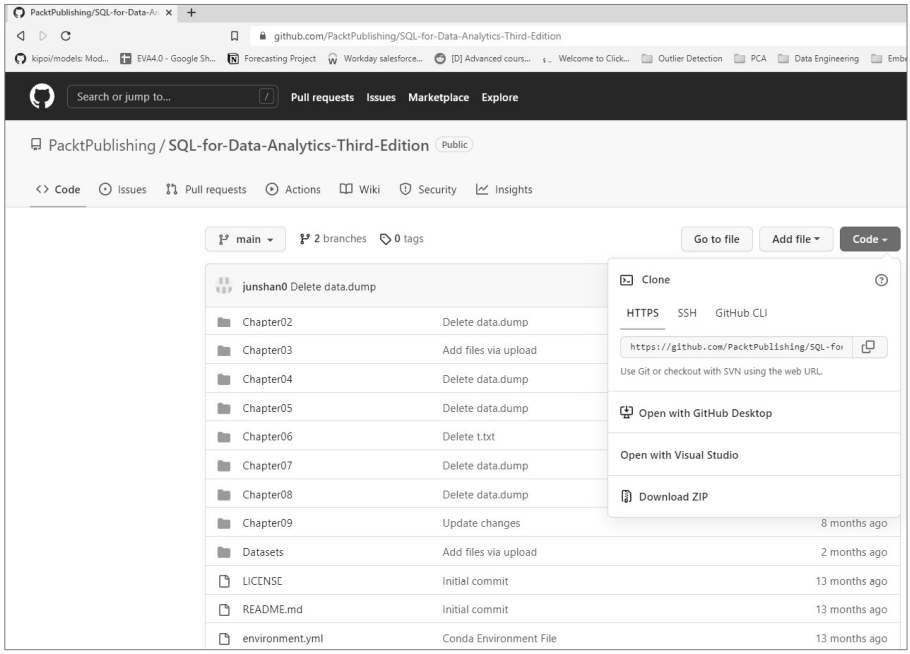
W tym ćwiczeniu użyjesz programu Microsoft Excel do utworzenia histogramu. Przyjmij, że jesteś analitykiem polityki zdrowotnej i chcesz zobaczyć rozkład wzrostu, aby dostrzec w nim wzorce. Aby wykonać to zadanie, musisz przygotować histogram.

Uwaga

Do tworzenia histogramów możesz użyć arkusza kalkulacyjnego, na przykład w programie Excel, albo języków do skryptowej analizy danych, takich jak Python. Dla wygody tu posłużysz się Excelem.

Wykonaj następujące kroki:

1. Wszystkie zbiory danych z tego rozdziału znajdziesz w serwisie GitHub. Aby uzyskać dostęp do plików, otwórz w dowolnej przeglądarce stronę *http://packt.link/hW355* (rysunek 1.3).



Rysunek 1.3. Pobieranie plików z kodem z serwisu GitHub

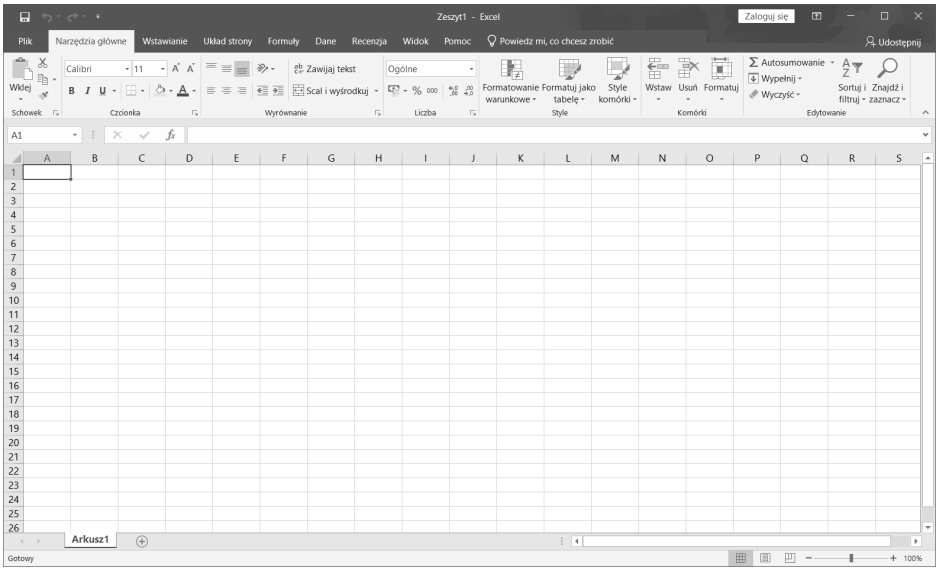
Kliknij menu rozwijane *Code* w prawym górnym rogu i wybierz opcję *Download ZIP*. W ten sposób pobierzesz plik zip zawierający cały kod z tej książki. Wypakuj zawartość tego pliku i przejdź do katalogu *Datasets*. Znajdziesz w nim pliki z danymi.

2. Otwórz pusty skoroszyt w programie Microsoft Excel (rysunek 1.4).

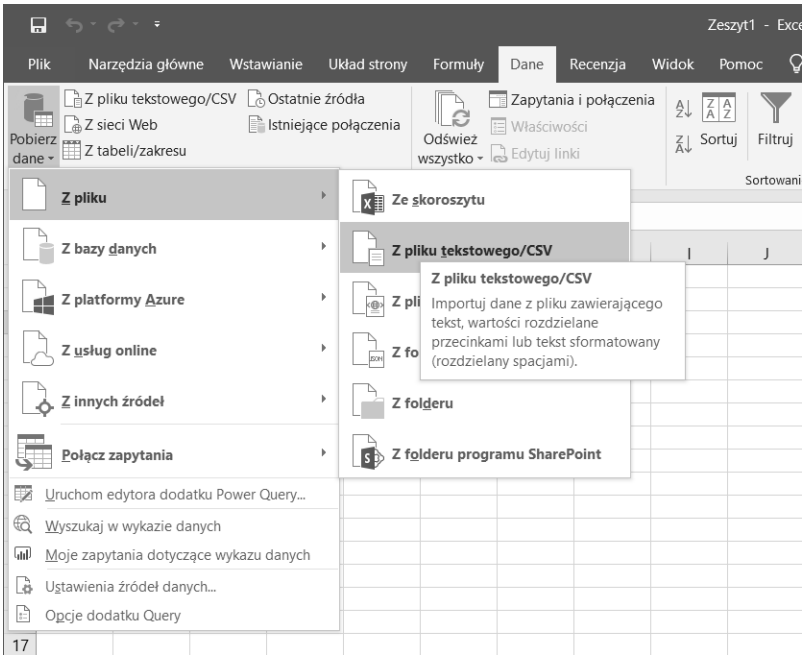
Uwaga

W polskiej wersji książki używany jest Excel w wersji 1808. Jeśli korzystasz z innej wersji Excela, ekran i menu mogą wyglądać inaczej, jednak proces przebiega tak samo i powinno być możliwe znalezienie menu i opcji potrzebnych w zadaniach z tej książki.

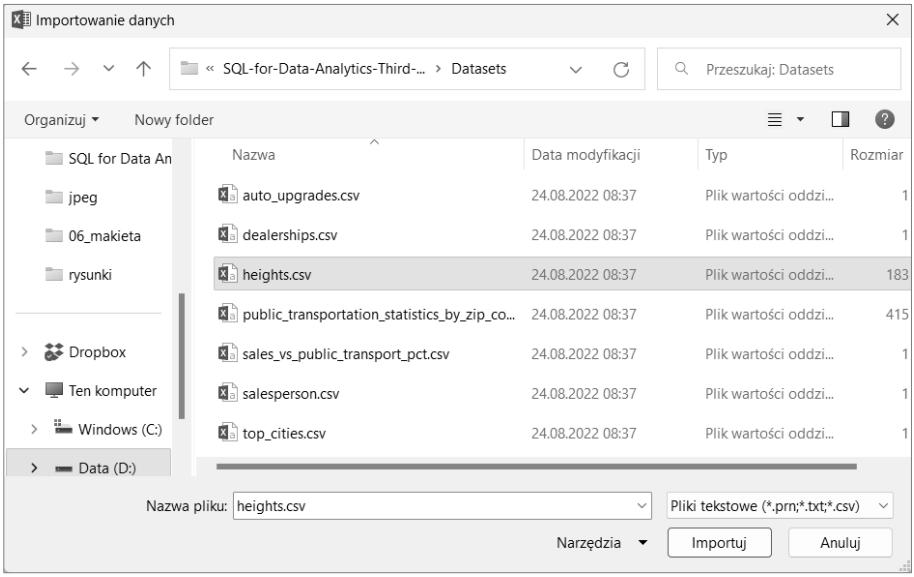
3. Przejdź do zakładki *Dane* i wybierz opcję *Pobierz dane/Z pliku/Z pliku tekstowego/CSV* (rysunek 1.5).
4. Znajdź plik ze zbiorem danych *heights.csv* w katalogu *Datasets* z repozytorium serwisu GitHub. Po wskazaniu tego pliku kliknij *Importuj* (rysunek 1.6).



Rysunek 1.4. Pusty skoroszyt w Excelu



Rysunek 1.5. Otwieranie pliku CSV



Rysunek 1.6. Wybieranie pliku heights.csv

5. Pojawi się okno dialogowe kreatora importu tekstu.

Uwaga

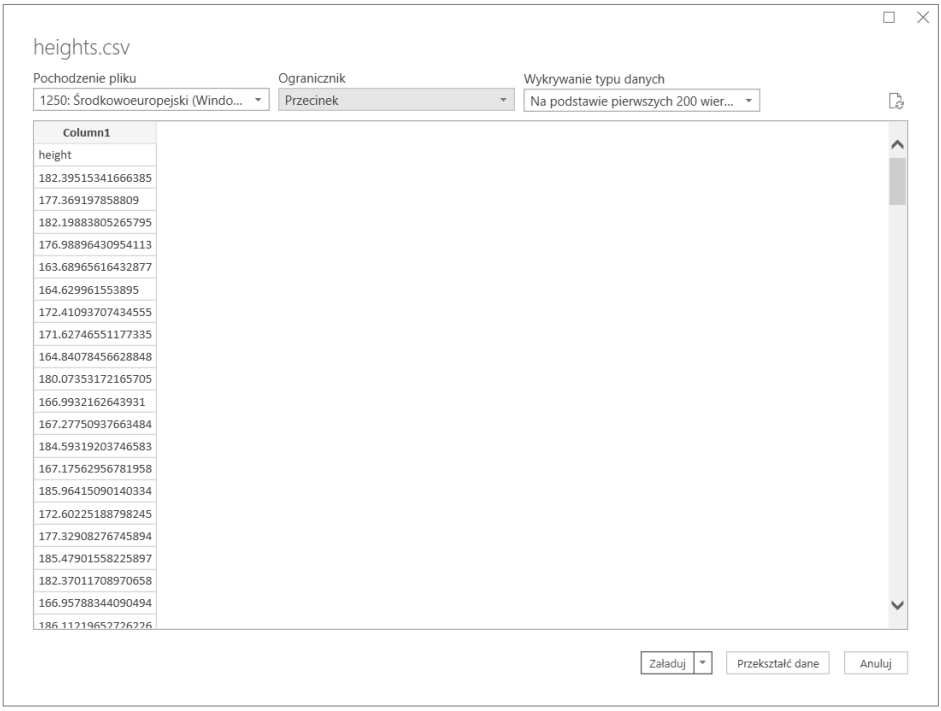
W górnej części okna widoczne jest menu rozwijane *Ogranicznik* (rysunek 1.7). **Ogranicznik** to token używany do rozdzielania różnych kolumn z tego samego wiersza. Jeśli masz w wierszu dwie kolumny, jedną z imieniem i drugą z wiekiem (na przykład `Sara` i `23`), potrzebujesz jakiegoś znaku do rozdzielenia tych dwóch wartości, aby komputery wiedziały, że należą one do innych kolumn. W plikach **CSV** (ang. *Comma-Separated Values*) jako ogranicznik zwyczajowo używany jest przecinek, dlatego wiersz wygląda tak:

```
Sara, 23
```

Plik *heights.csv* ma tylko jedną kolumnę, dlatego nie występują w nim ograniczniki. Możesz więc pozostawić opcję bez zmian. Teraz kliknij przycisk *Załaduj*[1].

6. 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 1.8.

[1] W zależności od ustawień systemu i programu Excel może być konieczna zmiana separatora części dziesiętnej z przecinka na kropkę lub zastąpienie w pliku kropek przecinkami — *przyp. tłum.*

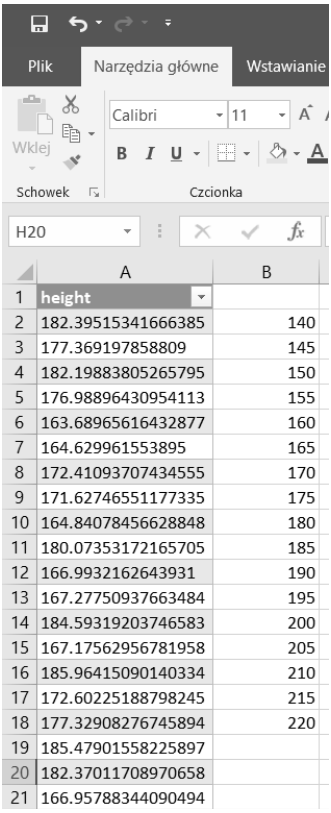


Rysunek 1.7. Wybieranie ogranicznika

7. 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).
8. W polu wyboru, które się pojawi, wybierz opcję *Histogram* i kliknij przycisk *OK*. Pojawi się okno dialogowe *Histogram*.
9. Aby podać opcję *Zakres komórek*, kliknij przycisk po prawej stronie pola tekstowego. Powinieneś wrócić do arkusza *Arkusz1*, gdzie widoczne będzie puste pole z przyciskiem ze strzałką (rysunek 1.9).

   Przeciągnij kursor i zaznacz wszystkie dane w arkuszu od komórki `A2` do `A10001`. Efekt jest pokazany na rysunku 1.10.

   Teraz kliknij przycisk ze strzałką, aby wrócić do okna *Histogram*.
10. Aby ustawić opcję *Zakres zbioru*, kliknij przycisk po prawej stronie pola tekstowego. Powinieneś wrócić do arkusza *Arkusz1*, gdzie widoczne będzie puste pole z przyciskiem ze strzałką. Przeciągnij kursor i zaznacz wszystkie dane w arkuszu od komórki `C2` do `C18`. Następnie kliknij przycisk ze strzałką.



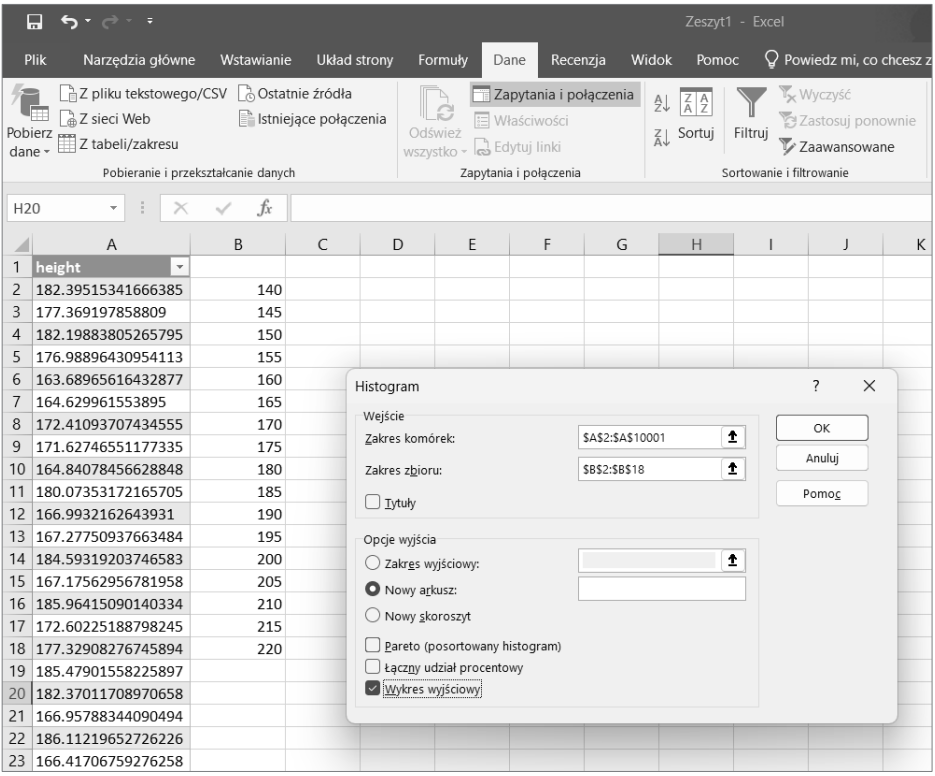
	A	B
1	height	
2	182.39515341666385	140
3	177.369197858809	145
4	182.19883805265795	150
5	176.98896430954113	155
6	163.68965616432877	160
7	164.629961553895	165
8	172.41093707434555	170
9	171.62746551177335	175
10	164.84078456628848	180
11	180.07353172165705	185
12	166.9932162643931	190
13	167.27750937663484	195
14	184.59319203746583	200
15	167.17562956781958	205
16	185.96415090140334	210
17	172.60225188798245	215
18	177.32908276745894	220
19	185.47901558225897	
20	182.37011708970658	
21	166.95788344090494	

Rysunek 1.8. Wprowadzanie danych w arkuszu Excela

Rysunek 1.9. Okno dialogowe do wprowadzania zakresu danych wejściowych

Rysunek 1.10. Wprowadzanie zakresu danych wejściowych

11. W sekcji *Opcje wyjścia* zaznacz pole *Nowy arkusz* i upewnij się, że zaznaczone jest pole *Wykres wyjściowy*, tak jak na rysunku 1.11. Następnie kliknij przycisk *OK*.



	A	B
1	height	
2	182.39515341666385	140
3	177.369197858809	145
4	182.19883805265795	150
5	176.98896430954113	155
6	163.68965616432877	160
7	164.629961553895	165
8	172.41093707434555	170
9	171.62746551177335	175
10	164.84078456628848	180
11	180.07353172165705	185
12	166.9932162643931	190
13	167.27750937663484	195
14	184.59319203746583	200
15	167.17562956781958	205
16	185.96415090140334	210
17	172.60225188798245	215
18	177.32908276745894	220
19	185.47901558225897	
20	182.37011708970658	
21	166.95788344090494	
22	186.11219652726226	
23	166.41706759276258	

Rysunek 1.11. Wybierz opcję Nowy arkusz

12. Jeśli zdecydujesz się zapisać histogram w nowym arkuszu, zostanie utworzony taki arkusz (zwykle o nazwie *Arkusz2*). Kliknij *Arkusz2*. Znajdź wykres i kliknij dwukrotnie tytuł *Histogram*. Zastąp go słowem *Wzrost*. Powinieneś uzyskać wykres podobny do tego z rysunku 1.12.

Rysunek 1.12. Rozkład wzrostu dorosłych mężczyzn

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. Jest to jeden z rozkładów, na które w praktyce będziesz natrafiać najczęściej. W tej książce nie omawiamy szczegółowo tego rozkładu, ale zwróć na niego uwagę w trakcie analiz danych, ponieważ często się pojawia.

Kwantyle

W poprzednim podpunkcie, „Rozkład danych", pokazaliśmy, jak obliczać rozkład, a także jak tworzyć jego wizualizację. Jednak dla każdego wzorca rozkładu można zbadać także inne aspekty. Na przykład gdy dane są dwa rozkłady normalne, jeden może być w większym stopniu skupiony wokół środka, przez co nachylenie krzywej jest większe, drugi rozkład natomiast może być bardziej rozciągnięty i wyglądać na bardziej wypłaszczony. Do zbadania cech każdego z tych rozkładów należy zastosować te same miary ilościowe.

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 (n – 1), które dzielą zmienną na cztery w przybliżeniu równe grupy wartości. Na rysunku 1.13 wymienione są nazwy kwantyli różnego rzędu.

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

Rysunek 1.13. Nazwy kwantyli rzędu N

Istnieją różne procedury obliczania kwantyli. Tu użyjesz jednej z najczęściej stosowanych technik obliczania kwantyli rzędu *N* dla jednej zmiennej:

1. Uporządkuj punkty danych od najmniejszego do największego według używanej zmiennej.
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.

Wyznacz indeks *i* dla *k*-tego punktu podziału. Użyj wzoru z rysunku 1.14.

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

Rysunek 1.14. Indeks

4. Jeśli *i* dla *k*-tego punktu podziału jest liczbą całkowitą, wybierz element o tym numerze spośród uporządkowanych punktów danych. Jeżeli *i* 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.
5. Powtarzaj kroki od 1. do 4. dla różnych wartości *k* do momentu obliczenia wszystkich punktów podziału.

Teraz, gdy znasz już kroki obliczania kwartyli, warto wykonać ćwiczenie, aby lepiej je zrozumieć.

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 — 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 (`Add-on Sales ($)`): 5000, 1700, 8200, 1500, 3300, 9000, 2000, 0, 0, 2300, 4700.

Uwaga

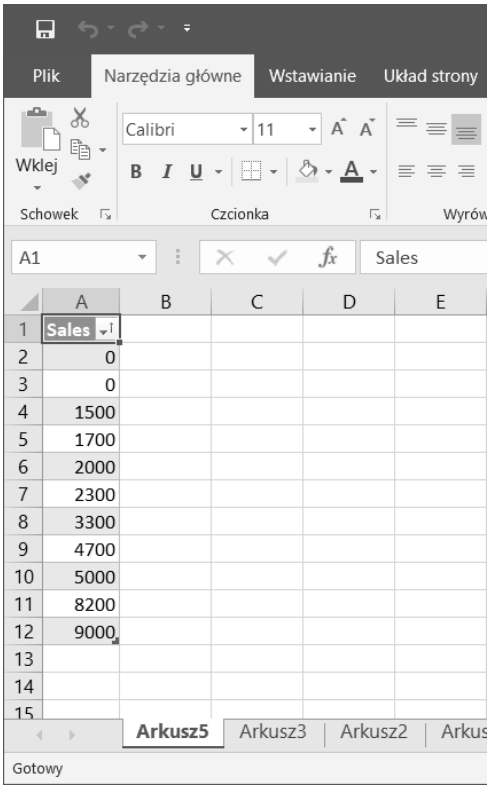
Wszystkie zbiory danych używane w tym rozdziale znajdziesz w serwisie GitHub: *https://packt.link/skue4*.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Otwórz pusty skoroszyt w programie Microsoft Excel.
2. Przejdź do zakładki *Dane* i wybierz opcję *Pobierz dane/Z pliku/Z pliku tekstowego/CSV*. Plik ze zbiorem danych *auto_upgrades.csv* znajdziesz

w katalogu *Datasets* w repozytorium w serwisie GitHub. Przejdź do tego pliku i kliknij przycisk *Importuj*[2].

3. 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. Na razie kliknij zakładkę z danymi z pliku *auto_upgrades.csv* i wybierz opcję sortowania.
4. Pojawi się okno dialogowe sortowania. Kliknij przycisk *Sortuj od najmniejszych do największych*. Wartości zostaną posortowane od najmniejszych do największych. Lista z rysunku 1.15 przedstawia posortowane wartości.



Rysunek 1.15. Posortowane wartości sprzedaży dodatkowego wyposażenia

5. Teraz ustal liczbę kwantyli rzędu *n* i punktów podziału, jakie musisz obliczyć. Kwartyle to kwantyle rzędu 4, zgodnie z rysunkiem 1.13. Ponieważ liczba punktów podziału jest o 1 mniejsza od liczby kwantyli rzędu *n*, wiadomo, że potrzebne są trzy punkty podziału.

[2] Podobnie jak w trakcie importowania poprzedniego pliku może zajść potrzeba zastąpienia w danych kropek przecinkami albo zmiany ustawień systemowych lub programu Excel — *przyp. tłum.*

6. 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 1.16 uzyskasz wynik *3,5*.

$$i = \left\lfloor \frac{k}{n}(d-1) \right\rfloor + 1$$
$$i = \left\lfloor \frac{1}{4}(11-1) \right\rfloor + 1$$
$$i = \frac{10}{4} + 1$$
$$i = \frac{10}{4} + 1$$
$$i = 2.5 + 1 = 3.5$$

Rysunek 1.16. Obliczanie indeksu pierwszego punktu podziału

7. 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.
8. 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. Kwartyle są przydatne, ponieważ dzielą zbiór danych na cztery podzbiory według kolejności. Na podstawie tych czterech podzbiorów łatwo jest ustalić połowę z większymi wartościami, połowę z mniejszymi wartościami, a także połowę wartości najbliższych medianie. Większość obecnie używanych narzędzi komputerowych, w tym SQL, umożliwia łatwe obliczanie kwantyli za pomocą wbudowanych mechanizmów, dlatego nie musisz robić tego ręcznie. Mimo to warto rozumieć, jak się to odbywa, czego można się dowiedzieć z przedstawionego przykładu.

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 1.1 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 (zobacz rysunek 1.12). 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 (na przykład miesiąc, w którym wygrywasz na loterii, w porównaniu z miesiącami, gdy otrzymujesz zwykłe wynagrodzenie) 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 tego zbioru danych niż średnia, która wynosi 38. Ta odporność na wartości odstające jest jednym z głównych powodów używania mediany.

Uwaga

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 (kolumna `Add-on Sales`). 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. Otwórz skoroszyt Excela i wpisz w kolumnie podane liczby.
2. Oblicz wartość modalną, aby wyznaczyć najczęściej występującą wartość. Ponieważ w tym zbiorze danych najczęściej pojawia się 0, wartość modalna to *0*.
3. 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.
4. Zaznacz cały zakres danych. W zakładce *Dane* wybierz opcję *AZ* z sekcji *Sortowanie i filtrowanie*. Oblicz medianę przez ustalenie środkowej wartości danych (rysunek 1.17).

   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.

Uwaga

Gdy porównasz średnią z medianą, zobaczysz, że znacznie się od siebie różnią. Wcześniej (w punkcie „Tendencja centralna") wspomnieliśmy, że jest to oznaka występowania wartości odstających w zbiorze danych. Należy wtedy ustalić, czy oczyścić dane przez usunięcie wartości odstających, czy tego nie robić. Dalej, w punkcie „Dyspersja", dowiesz się, jak stwierdzić, które wartości są odstające.

	A	B	C
1	0		
2	0		
3	1500		
4	1700		
5	2000		
6	2300		
7	3300		
8	4700		
9	5000		
10	8200		
11	9000		
12			
13			
14			
15			

Rysunek 1.17. Posortowane dane o sprzedaży dodatków

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

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. Na przykład w „Ćwiczeniu 1.03 — obliczanie miar tendencji centralnej dla sprzedaży dodatków" rozstęp wynosi 9000 (od 0 do 9000). 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.

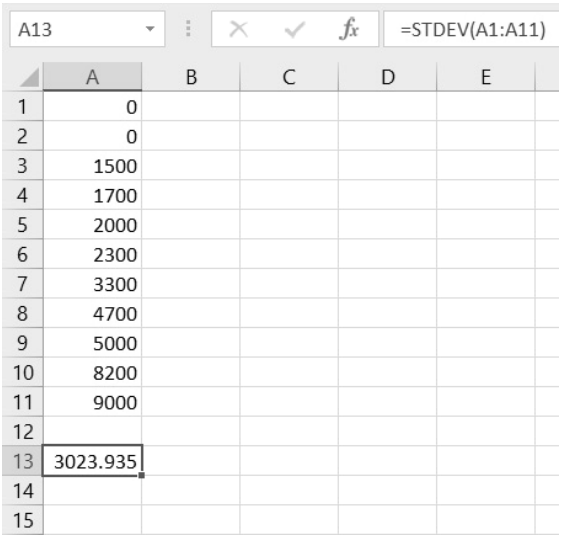
Warto zwrócić uwagę na to, że są dwa różne wzory na obliczanie odchylenia standardowego, co ilustruje rysunek 1.18. Gdy zbiór danych reprezentuje całą populację, należy obliczyć odchylenie standardowe dla populacji, używając wzoru A z tego rysunku.

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

Rysunek 1.18. Wzory na odchylenie standardowe dla populacji (A) i próbki (B)

Zmienna u_x oznacza tu średnią dla zbioru danych. Jeśli próbka reprezentuje tylko część obserwacji, należy posłużyć się wzorem B do obliczenia odchylenia standardowego dla próbki. Także tu zmienna u_x to średnia dla zbioru danych. 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.

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 przykład w „Ćwiczeniu 1.03 — obliczanie miar tendencji centralnej dla sprzedaży dodatków" możesz użyć funkcji `STDEV()` z Excela do obliczenia odchylenia standardowego dla próbki (rysunek 1.19).



Rysunek 1.19. Obliczanie odchylenia standardowego w Excelu

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).

> **Uwaga**
>
> 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 (Q3 + 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 rozstęp. 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. 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 1.20.
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.

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 1.20. Obliczanie sumy kwadratów różnic

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 punkcie zobaczysz, jak posłużyć się analizą dwuczynnikową do znajdowania wzorców.

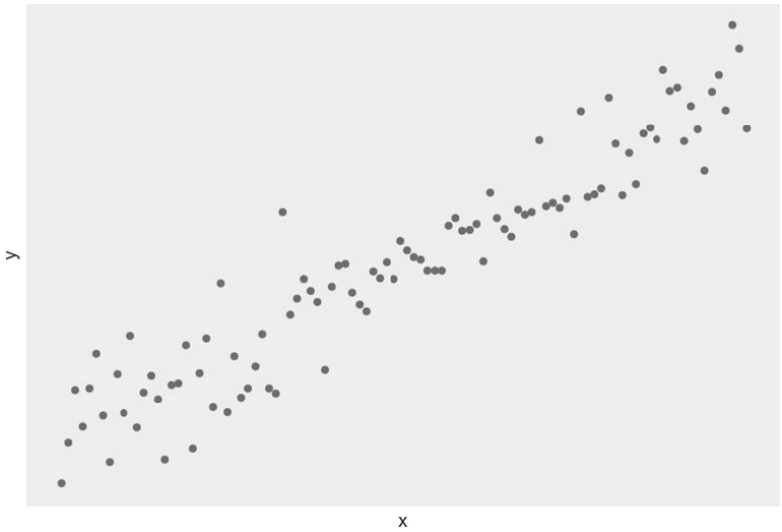
Analiza dwuczynnikowa

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

Wykresy punktowe

Jedną z najskuteczniejszych metod przeprowadzania analizy dwuczynnikowej jest stosowanie wykresów punktowych. 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, na przykład Excela.

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 1.21 ilustruje przykładowy trend liniowy.



Rysunek 1.21. Rosnący trend liniowy dla dwóch zmiennych — wieku i zarobków osób

Uwaga

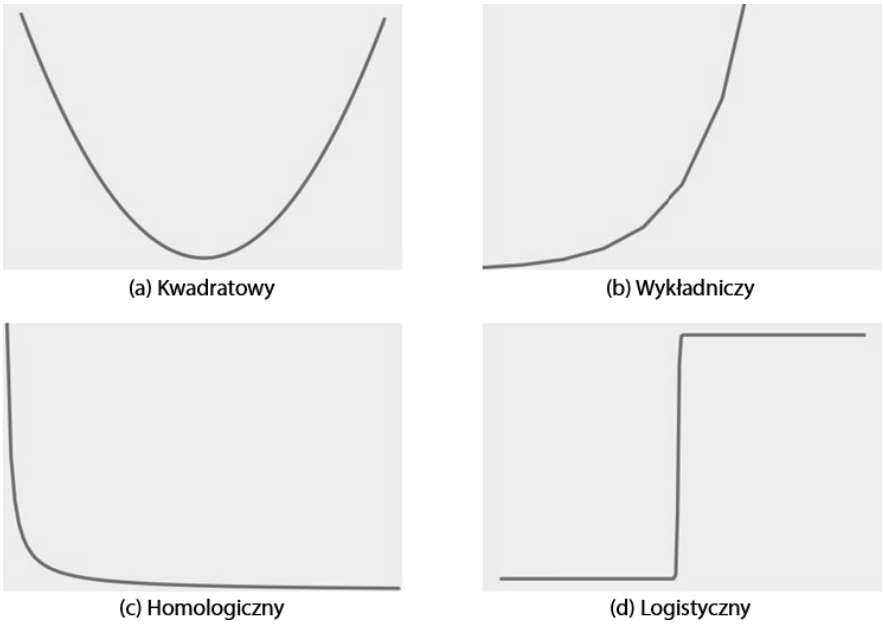
Wykresy punktowe są pomocne przede wszystkim w sytuacji, gdy liczba punktów jest niewielka (zwykle od 30 do 500). Jeśli liczba punktów jest duża i naniesienie ich na wykres kończy się uzyskaniem jednej wielkiej plamy, wybierz losową próbkę 200 punktów i utwórz wykres na ich podstawie, co może pomóc Ci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

Istnieje też wiele wartych uwagi trendów nieliniowych, w tym **kwadratowe**, **wykładnicze**, **homologiczne** i **logistyczne**. Na rysunku 1.22 pokazane są niektóre z nich.

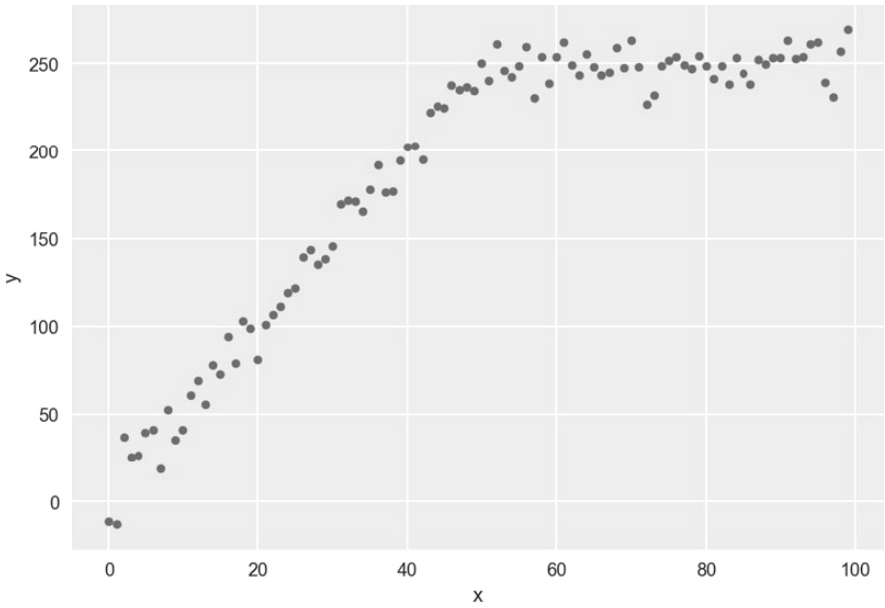
Uwaga

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.

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 1.23 przedstawia przykładową zmianę trendu. Tu trend liniowy załamuje się po punkcie **x=50**.

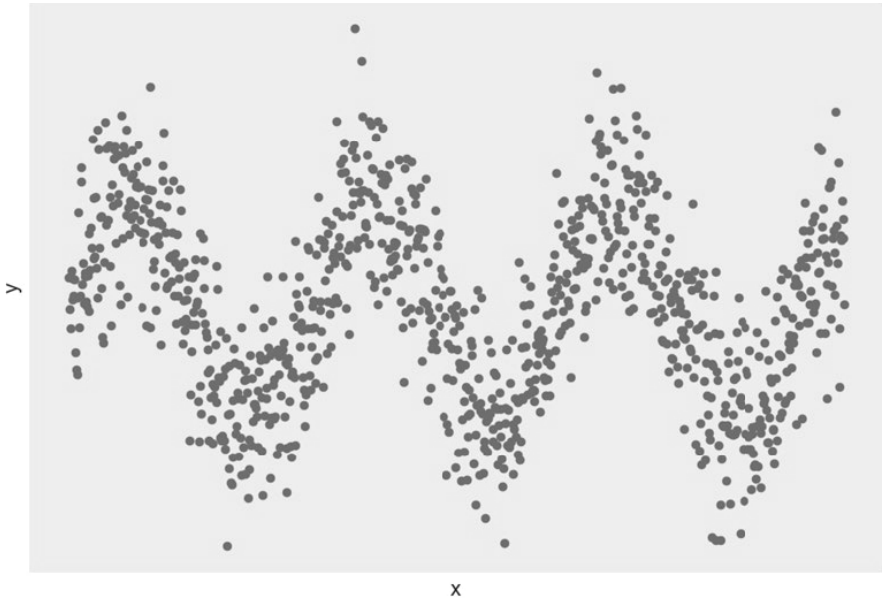


Rysunek 1.22. Inne często spotykane trendy



Rysunek 1.23. Przykład zmiany trendu

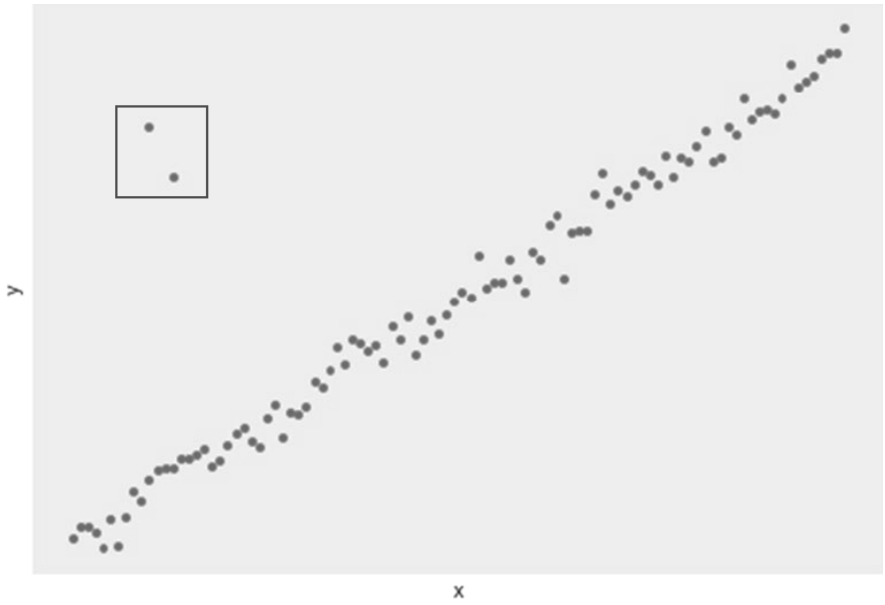
Innym wzorcem, na który często zwraca się uwagę, jest cykliczność, czyli powtarzające się wzorce w danych. Takie wzorce mogą wskazywać na to, że dwie zmienne zmieniają się cyklicznie, co może być przydatne w prognozowaniu. Bardzo często podawanym przykładem jest temperatura, która rośnie w ciągu dnia i spada w ciągu nocy. Rysunek 1.24 pokazuje przykład zmian cyklicznych.



Rysunek 1.24. 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 (tak jak dwa punkty w ramce w lewym górnym rogu na rysunku 1.25), 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.

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.



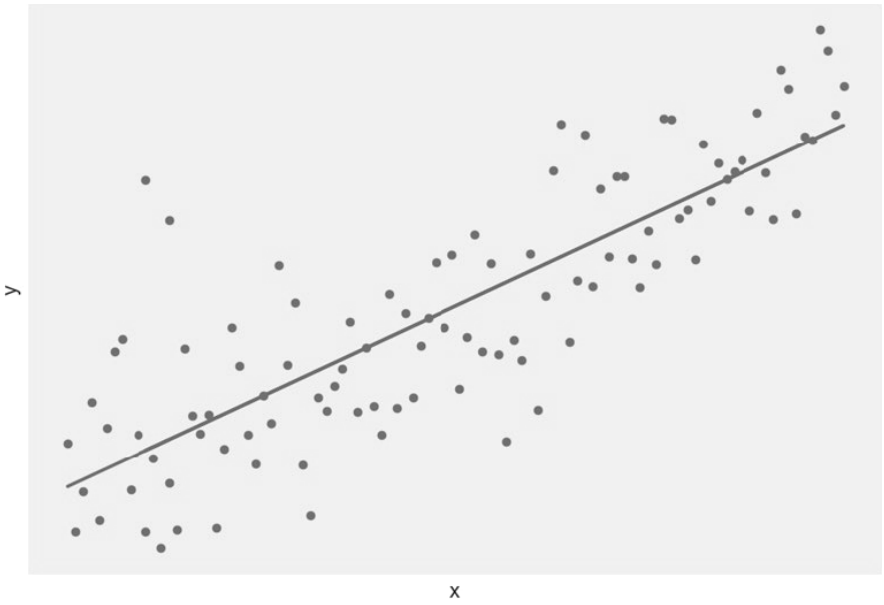
Rysunek 1.25. Wykres punktowy z dwiema wartościami odstającymi

Analiza trendów liniowych i współczynnik korelacji Pearsona

Jednym z najczęściej występujących trendów w trakcie analizy danych dwuczynnikowych jest trend liniowy. Trend liniowy określa, czy występuje relacja, w której wraz ze wzrostem wartości jednej zmiennej wartości innej zmiennej rosną lub maleją. Często się zdarza jednak, że niektóre trendy liniowe pasują do danych lepiej, a inne gorzej. Na rysunkach 1.26 i 1.27 znajdziesz przykładowe wykresy punktowe z linią najlepszego dopasowania. Ta linia jest obliczana **metodą najmniejszych kwadratów**.

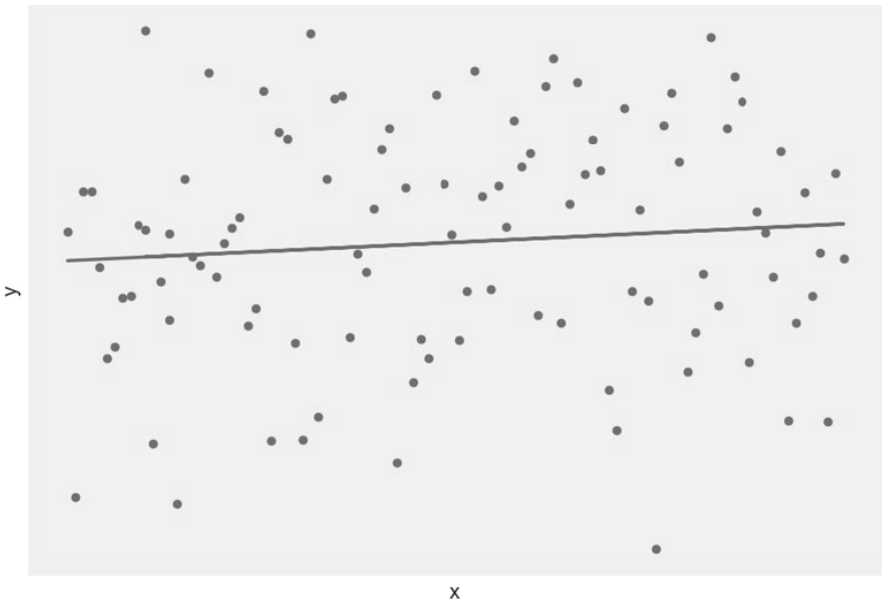
Uwaga

Choć omawianie metody najmniejszych kwadratów wykracza poza zakres tej książki, należy wiedzieć, że określa ona, jak dobrze dane dwuczynnikowe pasują do trendu liniowego. Jest to wartościowe narzędzie pomocne w zrozumieniu zależności między dwiema zmiennymi.



Rysunek 1.26. Wykres punktowy z wyraźnym trendem liniowym

Kolejny rysunek ilustruje wykres punktowy ze słabym trendem liniowym.



Rysunek 1.27. Wykres punktowy ze słabym trendem liniowym

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 1.28.

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

Rysunek 1.28. Wzór na obliczanie współczynnika korelacji Pearsona

W tym wzorze w mianowniku znajduje się iloczyn odchyleń standardowych zmiennych x i y, a w liczniku występuje kowariancja tych zmiennych. Te obliczenia są dość skomplikowane, dlatego prześledź przykład, aby przekształcić wzór na konkretne kroki.

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 1.29 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

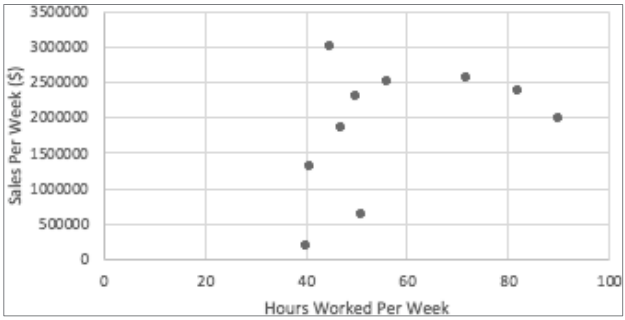
Rysunek 1.29. Dane 10 sprzedawców z salonu firmy ZoomZoom

Uwaga

Zbiór danych *salesperson.csv* potrzebny w tym ćwiczeniu możesz bezpośrednio pobrać z serwisu GitHub. Oto odsyłacz do katalogu *Datasets* — *https://packt.link/mriXZ*.

Oto kroki niezbędne do wykonania tego ćwiczenia:

1. Najpierw utwórz wykres dwóch zmiennych w Excelu na podstawie danych z tego scenariusza (rysunek 1.30). Pomoże Ci to na ogólnym poziomie ocenić, jakiego współczynnika korelacji Pearsona możesz oczekiwać.



Rysunek 1.30. 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.

2. Teraz oblicz średnią każdej zmiennej. Powinieneś uzyskać 57,40 dla zmiennej `Hours Worked Per Week` i 1 861 987,4 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 1.31.

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	179,480.58	-17.40	302.76	-1,682,506.85	2,830,829,303,631.31	29,275,619.21
56	2,495,037.73	-1.40	1.96	633,050.29	400,752,674,381.30	-886,270.41
50	2,285,369.51	-7.40	54.76	423,382.07	179,252,379,435.48	-3,133,027.34
82	2,367,896.33	24.60	605.16	505,908.90	255,943,812,657.79	12,445,358.88
41	1,309,745.16	-16.40	268.96	-552,242.27	304,971,527,314.18	9,056,773.27
51	623,013.69	-6.40	40.96	-1,238,973.75	1,535,055,945,620.25	7,929,431.98
45	2,989,943.37	-12.40	153.76	1,127,955.94	1,272,284,593,638.99	-13,986,653.61
90	1,970,316.24	32.60	1,062.76	108,328.81	11,735,131,115.82	3,531,519.21
47	1,845,840.39	-10.40	108.16	-16,147.04	260,726,862.48	167,929.20
72	2,553,231.33	14.60	213.16	691,243.90	477,818,127,736.76	10,092,160.92

Rysunek 1.31. Obliczenia współczynnika korelacji Pearsona

4. Oblicz sumy kwadratów i sumę iloczynów różnic. Powinieneś otrzymać 2812,40 dla zmiennej `Hours Worked Per Week` (x), 7 268 904 226 420,96 dla zmiennej `Sales Per Week ($)` (y) i 54 492 841,19 dla iloczynu różnic.
5. Oblicz pierwiastki kwadratowe sum różnic. Powinieneś uzyskać 53,03 dla zmiennej `Hours Worked Per Week` (x) i 2 696 090,55 dla zmiennej `Sales Per Week ($)` (y).

- Podstaw te wartości do wzoru z rysunku 1.32. Otrzymasz wynik *0,38*. Obliczenia są pokazane na rysunku 1.32: 54492841,19/(53,03 * 2696090,55) = 0,38.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{54492841{,}19}{(53{,}03) * (2696090{,}55)} \approx 0.38$$

Rysunek 1.32. 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 3., „Przygotowywanie danych za pomocą SQL-a", zobaczysz, że współczynnik korelacji Pearsona można obliczyć za pomocą SQL-a.

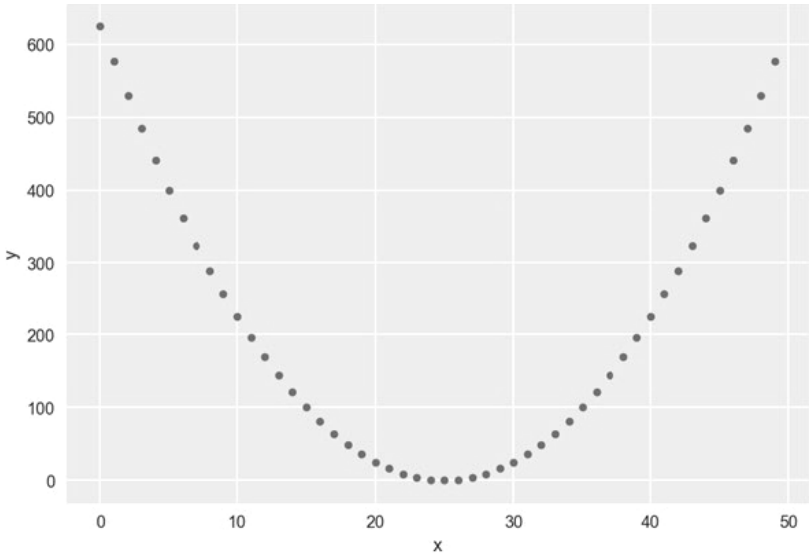
Aby zinterpretować współczynnik korelacji Pearsona, porównaj uzyskaną wartość z tabelą z rysunku 1.33. Im wynik jest bliższy zeru, tym korelacja jest słabsza. Im wyższa wartość bezwzględna współczynnika korelacji Pearsona, tym bardziej prawdopodobne jest, że punkty pasują do linii prostej.

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 1.33. 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.

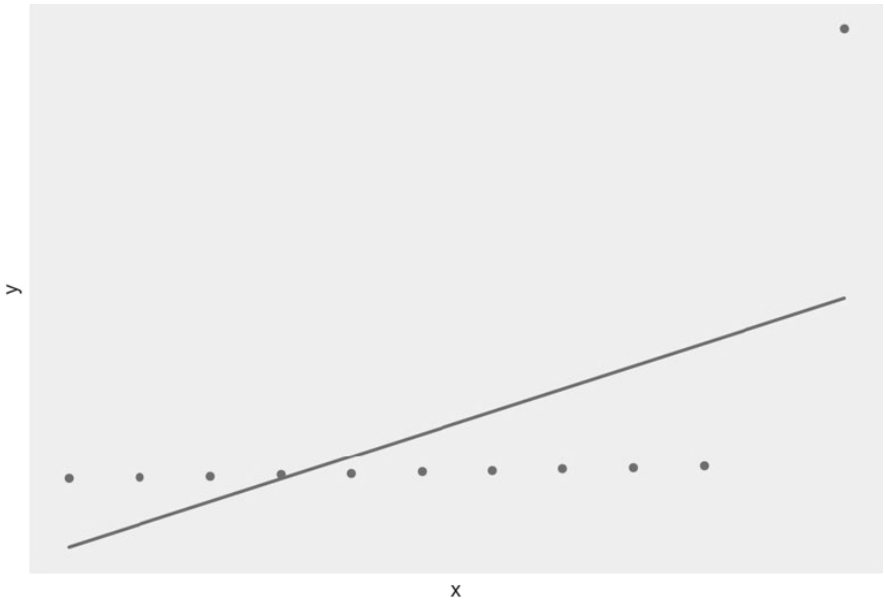
Przyjrzyj się na przykład punktom z rysunku 1.34. Jeśli obliczysz współczynnik dla tych dwóch zmiennych, otrzymasz wynik –0,08. Jednak krzywa wskazuje na bardzo silną zależność kwadratową. Dlatego gdy sprawdzasz współczynniki korelacji danych dwuczynnikowych, pamiętaj, że zależność między dwiema zmiennymi może być nieliniowa.



Rysunek 1.34. Silna nieliniowa zależność o niskim współczynniku korelacji

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 1.35. 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 liczba punktów danych jest mała.



Rysunek 1.35. 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. Wprawdzie możliwe jest, że większa liczba godzin pracy skutkuje wyższą sprzedażą, jednak wysoki współczynnik korelacji nie wystarczy jako dowód tej tezy.

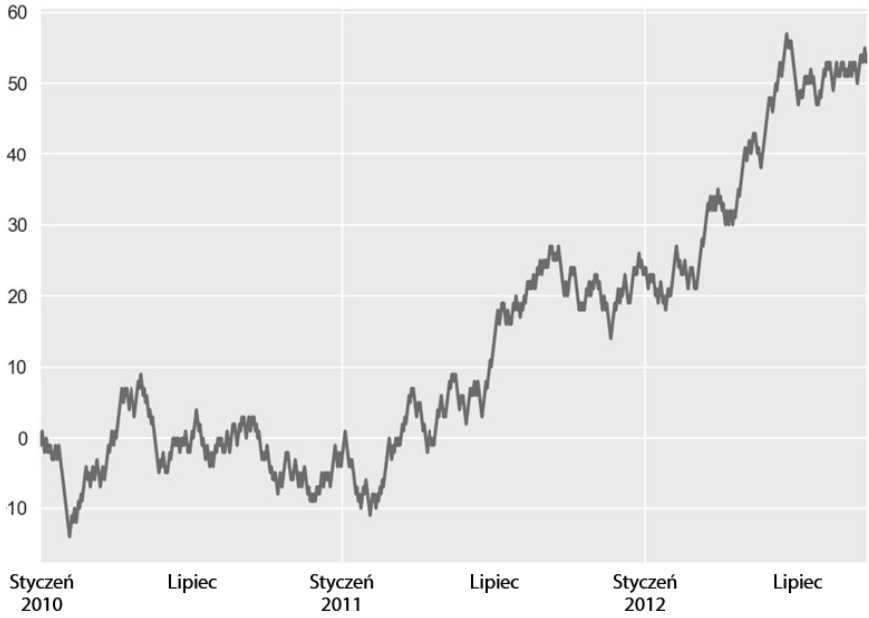
Możliwe nawet, że związek przyczynowy działa w drugą stronę: większa liczba transakcji wymaga więcej pracy papierkowej, co przekłada się na większą liczbę godzin w biurze. W tym scenariuszu większa liczba godzin nie powoduje wyższej sprzedaży.

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 1.36. Widoczny jest tam szereg czasowy obejmujący okres od stycznia 2010 roku do końca 2012 roku.



Rysunek 1.36. Przykładowy szereg czasowy

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.

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 z wykorzystaniem statystyki. 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.

To zadanie uczy, jak radzić sobie z danymi, procesami i typami danych. W całym podrozdziale pokazaliśmy, jak stosować techniki jedno- i dwuczynnikowej analizy danych. Jak jednak postępować z niekompletnymi danymi? Następny punkt pomoże Ci zrozumieć, co robić w takim scenariuszu.

Uwaga

Rozwiązanie tego zadania znajdziesz w „Dodatku".

Praca z niepełnymi danymi

We wszystkich dotychczasowych przykładach zbiory danych były oczyszczone i łatwe do zrozumienia. Jednak w praktyce zbiory danych są bardziej skomplikowane. Jednym z wielu problemów, z jakimi trzeba sobie radzić w trakcie pracy z danymi, są brakujące wartości.

Szczegóły przygotowywania danych omówiliśmy w rozdziale 3., „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.

- **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 z danologiem, 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.

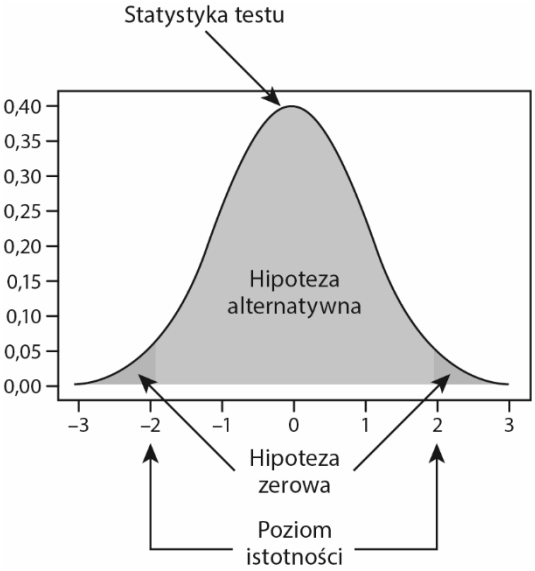
Przekonasz się, że analiza danych często jest bardziej sztuką niż nauką. Dotyczy to między innymi pracy z niekompletnymi danymi. Gdy nabierzesz doświadczenia, będziesz znać kombinacje strategii, które sprawdzają się w różnych scenariuszach.

Testy istotności statystycznej

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? Aby to określić, potrzebna jest metoda oparta na statystyce.

Testy istotności statystycznej to techniki sprawdzania, czy posiadane dane potwierdzają określone hipotezy. W testach istotności statystycznej trzeba uwzględnić kilka ważnych elementów (rysunek 1.37). 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.



Rysunek 1.37. Elementy testów istotności statystycznej

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.

Często używane testy istotności statystycznej

Testy istotności statystycznej są ważnym elementem analizy danych. W typowym scenariuszu analityk zbiera dane z rzeczywistego świata i tworzy modele pasujące do tych danych. Jednak jak trafne są te modele? Czy na ich podstawie można precyzyjnie prognozować, co stanie się w rzeczywistym świecie? Aby uzyskać odpowiedź na te pytania, trzeba przeprowadzić testy istotności statystycznej.

Wszystkie testy istotności statystycznej mają cztery aspekty opisane w poprzednim punkcie. W poszczególnych testach istotności statystycznej stosuje się różne sposoby obliczania tych aspektów. 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.

SQL i analityka

W tym rozdziale omówiliśmy różne techniki stosowane w analityce danych. Wszystkie te techniki wymagają składowania i przetwarzania dużych ilości danych. Choć obecnie na rynku dostępnych jest wiele narzędzi pomocnych w wykonywaniu tych zadań, najważniejszym z nich są relacyjne bazy danych.

Stosowanie relacyjnej bazy danych to wygodny i łatwy do zrozumienia sposób składowania zbiorów danych. Nowoczesne systemy zarządzania relacyjnymi bazami danych, na przykład PostgreSQL, zawierają zaawansowane narzędzie do przetwarzania i analizowania danych. Tym narzędziem jest SQL. Za pomocą SQL-a możesz oczyszczać dane, przekształcać je na bardziej przydatny format i analizować za pomocą metod statystycznych, aby wykrywać interesujące wzorce. Reszta książki pozwoli Ci zrozumieć, jak stosować SQL do tych celów w produktywny i wydajny sposób.

Podsumowanie

Analityka danych to zaawansowana metoda analizowania surowych danych w celu wykrywania wzorców i generowania prognoz, które pomagają zrozumieć świat. Ostatecznym celem analityki 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.

Choć statystyka jest bardzo ważna, wymaga dostępu do dużych ilości danych. Zarówno składowanie danych, jak i obliczenia z ich wykorzystaniem mogą być niezwykle wymagające. Opracowano różne narzędzia, by wykorzystać możliwości komputerów do obliczeń statystycznych. Jednymi z najważniejszych narzędzi tego rodzaju są relacyjne bazy danych i SQL. Reszta książki zawiera omówienie aspektów i stosowania SQL-a. Zaczynamy od następnego rozdziału, w którym przedstawiamy wprowadzenie do relacyjnych baz danych i SQL-a. Dowiesz się też, jak tworzyć, wczytywać, aktualizować i usuwać (ang. *create*, *read*, *update*, *delete* — CRUD) zbiory danych.

Skorowidz

A

ACID, atomicity, consistency, isolation, durability, 80
aktualizowanie tabel, 111, 115
Anaconda Navigator, 230
analityka danych, 38
analiza
  danych, 159, 188
    z szeregów czasowych, 259
  dwuczynnikowa, 40, 58
    analiza trendów liniowych, 62
    interpretowanie współczynnika korelacji Pearsona, 66
    szeregi czasowe, 69
    współczynnik korelacji Pearsona, 62
    wykresy punktowe, 58
  jednoczynnikowa, 40
    dyspersja, 55
    kwantyle, 49
    miary tendencji centralnej, 52
    rozkład danych, 41
  post hoc, 302
  sekwencji, 271
  tekstu, 286, 288
  trendów liniowych, 62
analizy geoprzestrzenne, 261, 265

B

bazy danych
  metody skanowania, 303
B-drzewo, 311
biblioteka psycopg2, 230

C

CRUD, create, read, update, delete, 79

D

dane
  częściowo ustrukturyzowane, 77
  ilościowe, 38
  jakościowe, 38
  niepełne, 70
  nieustrukturyzowane, 77
  ustrukturyzowane, 77
data i godzina, 106, 252
długość i szerokość geograficzna, 262
dyspersja, 55

E

edytor kwerend, 85
eksportowanie
  danych, 212, 402
  danych do pliku, 223
ETL, extract, transform, load, 184
Excel
  odczyt plików CSV, 223

F

format
  CSV, 216, 223
  JSON, 106, 274, 277
  JSONB, 280
    dodawanie i usuwanie elementów, 282
    modyfikowanie danych, 282
    przeszukiwanie obiektów, 283
funkcja, 328
  ARRAY_AGG, 269
  CASE WHEN, 145, 147
  CASTING, 152
  COALESCE, 148
  COUNT DISTINCT, 163
  COUNT(*), 161, 163, 191–196
  COUNT, 164
  DATE_PART, 256
  DATE_TRUNC, 256
  DENSE_RANK, 202, 204
  DISTINCT, 153
  DISTINCT ON, 153
  EXTRACT, 255
  GREATEST, 151
  JSONB_ARRAY_ELEMENTS, 285
  JSONB_PATH_EXISTS, 280
  JSONB_PATH_QUERY, 281
  JSONB_PATH_QUERY_ARRAY, 282
  JSONB_PRETTY, 279, 285
  LAG, 202
  LEAD, 202
  LEAST, 151
  NOW, 254
  NTILE, 202
  NULLIF, 149
  RANK, 202, 203
  REGEXP_REPLACE, 290
  ROW_NUMBER, 202
  ROW_TO_JSON, 275
  STRING_TO_ARRAY, 289
  TO_CHAR, 256, 257
  TO_TSQUERY, 294
  TO_TSVECTOR, 293, 294
  TS_LEXIZE, 290
  UNNEST, 289
funkcje
  agregujące, 160, 393
    analiza danych, 159, 165, 185

funkcje
dla zbiorów uporządkowanych, 177
do oczyszczania danych, 181
sprawdzanie unikatowości danych, 184
z klauzulą GROUP BY, 166
z klauzulą HAVING, 178
nie przyjmujące argumentów, 329
okna, 191, 209, 399
analiza danych, 189
obliczanie statystyk, 202
przyjmujące argumenty, 333, 335, 420
tablicowe, 271

H

hasło, 247
haszowanie, 318
hipoteza
alternatywna, 72
zerowa, 72
histogram, 42, 386

I

importowanie danych, 212, 402
indeks
bazy danych, 310
GIN, 296
w postaci B-drzewa, 311
z haszowaniem, 318, 320, 323, 418
indeksy
skuteczne wykorzystanie, 324
instrukcja
ALTER, 111
CASE WHEN, 145, 147
COPY, 213–216, 221, 222, 243
CREATE, 107
CREATE TABLE, 108, 109
CREATE VIEW, 220
DROP TABLE, 117
EXPLAIN, 304–307
EXPLAIN ANALYZE, 316
INSERT, 113
LIMIT, 224
SELECT, 83, 86, 99, 109, 387
UPDATE, 114, 115
integralność referencyjna, 124
interfejs narzędzia psql, 224
interpreter Pythona, 229
interpretowanie współczynnika korelacji Pearsona, 66
istotność statystyczna, 71

J

język
JSONPath, 280
Python, 228
SQL, 37, 75
JSON, JavaScript Object Notation, 106, 274
JSONB, 277
JSONPath, 280
Jupyter Notebook, 231, 235, 237

K

klasyfikowanie nowego zbioru danych, 39
klauzula
AND, 91
COUNT DISTINCT, 184
CROSS JOIN, 136
DISTINCT ON, 266
FULL OUTER JOIN, 135
GROUP BY, 166, 167, 173, 181
GROUPING SETS, 175, 185
HAVING, 178, 180
IN, 93
IS NOT NULL, 98
IS NULL, 98
JOIN, 124
LEFT OUTER JOIN, 132
LIMIT, 97
NOT IN, 93
OR, 91
ORDER BY, 94, 269, 272
PARTITION BY, 191–193, 196
UNION, 141, 142
UNION ALL, 141
WHERE, 90
WINDOW, 200
WITH, 144
konwerter obiektowo-relacyjny, ORM, 233
kwantyle, 49
rzędu N, 49
kwerendy
optymalizacja, 302
plany wykonywania, 304, 309, 415
zakończenie działania, 325
zwiększanie wydajności, 310, 317, 320

Ł

łączenie tabel, 124

M

mediana, 53
metoda najmniejszych kwadratów, 62
modelowanie atrybucji, 273

N

narzędzie
Anaconda Navigator, 230
Jupyter Notebook, 231, 235, 237
pandas, 233, 235, 237
pgAdmin, 83, 228
pgAdmin 4, 197
psql, 214, 224, 228
SQLAlchemy, 233

O

obiekty zagnieżdżone, 276
obliczanie
dyspersji, 57
kwartyli, 50
miar tendencji centralnej, 54
odchylenia standardowego, 56

statystyk, 202
sumy kwadratów różnic, 58
współczynnika korelacji Pearsona, 64–66
odchylenie standardowe, 55
ograniczenia kolumn, 107
okno, 189
operacje CRUD, 86
operator
#>, 277
?, 281
@@, 295
||, 297
=, 318
->, 277
operatory logiczne, 294
ORM, Object-Relational Mapper, 233

P

pakiet
pandas, 233, 235, 237
SQLAlchemy, 235
pgAdmin, 83, 228
edytor kwerend SQL-owych, 85
interfejs programu, 83
planer kwerend, 304, 305
plik .pgpass, 247
pliki CSV, 44
podkwerendy, 139
polecenie, *Patrz także* instrukcja
\copy, 218, 221, 222, 225
\df, 334
\sf, 335
pg_cancel_backend, 325
pg_sleep, 325, 326
pg_terminate_backend, 326
psql, 214, 224
połączenie z bazą, 214, 247
PostgreSQL, 228
analiza sekwencji, 271
analiza tekstu, 286, 288
długość i szerokość geograficzna, 262
funkcje, 329
funkcje tablicowe, 271
indeks w postaci B-drzewa, 311
indeks z haszowaniem, 318
JSONB, 277
kończenie pracy kwerend, 325
optymalizowanie wyszukiwania tekstu, 296
planer kwerend, 304
przekształcanie typów danych, 255
stosowanie formatu JSON, 274
strategie indeksowania, 310
tablice, 268
tokenizacja tekstu, 286
wyszukiwanie tekstu, 293
wyświetlanie dat, 254
zarządzanie relacyjnymi bazami danych, 81
poziom istotności, 72
predykat złączenia, 127
psql, 214, 224, 228
punkty podziału, 49
Python, 228
dostęp do baz PostgreSQL, 233
odczyt plików CSV, 245
pakiet pandas, 233, 235, 237
pakiet SQLAlchemy, 235
pobieranie danych z bazy, 237
wizualizowanie danych, 238
zapis danych w bazie, 237, 238
zapis plików CSV, 245
zwiększanie szybkości zapisu, 243

R

ramka
danych, DataFrame, 233, 237
okna, 204, 209
relacja, 78
relacyjne bazy danych, 78
system zarządzania, 81
rozkład
bezwzględnej częstości występowania, 41
danych, 41
względnej częstości występowania, 42
rozstęp, 55
ćwiartkowy, 57

S

skanowanie
baz danych, 303
indeksu, 310, 312, 317, 416
sekwencyjne, 306
sekwencyjne równoległe, 308
skośny zbiór danych, 53
słowo kluczowe
FUNCTION, 328
OVER, 191, 202
PARTITION BY, 191–193, 196
PRECEDING, 208
RANGE, 205
ROWS, 205
TEMP, 220
TRIGGER, 337
WINDOW, 200
WITH, 216
SQL, Structured Query Language, 37, 75
funkcje, 328
wydajność kodu, 301
SQLAlchemy, 233
statystyki, 38, 202
opisowe, 40
wieloczynnikowe, 40
struktury danych, 106
studium przypadku, 347
analiza
czasu rozpoczęcia sprzedaży, 358
skuteczności kampanii e-mailowej, 377
wzrostu sprzedaży, 368
analizowanie hipotezy, 425
badania terenowe, 382
ilościowa ocena spadku sprzedaży, 422

studium przypadku
metoda naukowa, 347
pobieranie informacji sprzedażowych, 351
wnioski, 382
wstępne zbieranie danych, 348
sumy, 140
surowe dane, 38
system zarządzania
bazami danych, SZBD, 228, 303
relacyjnymi bazami danych, SZRBD, 81
szeregi czasowe, 69, 259

Ś

średnia, 53
krocząca, 205

T

tabele
aktualizowanie, 111, 115
istniejących wierszy, 114
dodawanie
danych, 112
kolumn, 111
modyfikowanie, 120
tworzenie, 108, 109, 120
usuwanie, 116–119
danych, 116
kolumn, 111
wartości z wiersza, 116
wierszy, 117
złączenia, 124
tablice, 106
w PostgreSQL, 268
techniki
analizy dwuczynnikowej, 41
analizy jednoczynnikowej, 41
tendencja centralna, 52
test
A/B, 71
Pearsona, 73
T dla dwóch próbek, 72
Z dla dwóch próbek, 72
zgodności chi-kwadrat, 73
testy istotności statystycznej, 71, 72
tokenizacja tekstu, 286
trend liniowy, 59, 63
trendy nieliniowe, 60
tworzenie
funkcji, 328, 329, 332–335
histogramu, 42
indeksów, 312
indeksów z haszowaniem, 320
tabel, 103, 108, 109
tymczasowych widoków, 219, 225
wyzwalaczy, 337, 339
typ danych
DATE, 252, 254, 259
INTERVAL, 257
JSONB, 277
logiczny, 105
POINT, 264
TIMESTAMP, 254
tsquery, 294
tsvector, 295
typy danych
liczbowe, 104
w SQL-u, 104
złożone, 251, 411
znakowe, 104

U

usuwanie tabel, 116–119

W

wariancja, 55
wartość
modalna, 52
NULL, 116
odstająca, 53
widoki tymczasowe, 219, 225
wnioskowanie statystyczne, 40
współczynnik korelacji Pearsona, 62, 64, 66
interpretowanie, 66
wykres punktowy, 58, 62
wyrażenie
ILIKE, 292, 293
WITH, 144
wyszukiwanie
tekstu, 293
tekstu zoptymalizowane, 296
wyzwalacz, 336–339
do aktualizowania pola, 339
INSTEAD OF, 336

Z

zarządzanie relacyjnymi bazami danych, 81
złączenia
krzyżowe, 136
wewnętrzne, 127
zewnętrzne, 131
lewostronne, 131
pełne, 135
prawostronne, 133

SQL: SPÓJRZ NA DANE OKIEM WYSZKOLONEGO ANALITYKA!

Język SQL zwykle służy do pracy z bazami danych, jednak można go używać również do wydajnego przetwarzania ich wielkich zbiorów. W tym celu trzeba dobrze poznać to narzędzie. Wysiłek włożony w zrozumienie SQL-a na pewno się opłaci — dzięki analizie danych można wydobywać z nich bezcenną wiedzę, która bezpośrednio przekłada się na zyski firmy.

Ta książka stanowi świetne wprowadzenie do analizy danych. Dzięki niej nauczysz się korzystać z surowych danych, nawet jeśli nie masz odpowiedniego doświadczenia. Zaczniesz od formułowania hipotez i generowania statystyk opisowych, a następnie przystąpisz do pisania zapytań w języku SQL w celu agregowania, przeliczania i łączenia danych z różnych zbiorów. Zapoznasz się też z zaawansowanymi technikami, takimi jak analiza geoprzestrzenna i analiza tekstu. W książce omówiono również profilowanie i automatyzację, które umożliwiają szybsze i wydajniejsze pobieranie informacji. To wszystko pozwoli Ci na skuteczne korzystanie z SQL-a w codziennych scenariuszach biznesowych.

NAJCIEKAWSZE ZAGADNIENIA:

- gruntowne wprowadzenie do analityki danych
- przygotowywanie danych do analizy
- optymalizacja kwerend i złożone typy danych
- funkcje agregujące, funkcja okna i inne metody analizy danych w SQL
- jak odkrywać prawdę za pomocą SQL-a

Jun Shan od ponad 20 lat projektuje systemy zarządzania danymi. Jest architektem rozwiązań chmurowych.

Matt Goldwasser jest kierownikiem do spraw danologii stosowanej w T. Rowe Price NYC Technology Development Center.

Upom Malik kieruje pracami w dziedzinie danologii i analityki. Używa SQL-a do rozwiązywania różnorodnych problemów.

Benjamin Johnston od ponad 10 lat projektuje i rozwija urządzenia medyczne.

Helion
helion.pl
HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-0173-5
9 788328 901735
Cena: 109,00 zł
Packt>