O'REILLY®

Analiza danych z wykorzystaniem SQL-a

Zaawansowane techniki przekształcania danych we wnioski

Cathy Tanimura

Helion

Tytuł oryginału: SQL for Data Analysis: Advanced Techniques for Transforming Data into Insights

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-283-8895-6

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

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

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

Printed in Poland.

Spis treści

ROZDZIAŁ 6.

Wykrywanie anomalii

Anomalia to element, który różni się od pozostałych elementów z tej samej grupy. W danych anomaliami są rekordy, obserwacje lub wartości różniące się od pozostałych punktów danych w sposób, który rodzi wątpliwości lub podejrzenia. Anomalie mają wiele różnych nazw, w tym *wartości odstające*, *szum*, *odchylenia* i *wyjątki*. W tym rozdziale używam wymiennie określeń *anomalia* i *wartość odstająca*. W innych tekstach możesz zetknąć się z jeszcze innymi słowami. Wykrywanie anomalii może być ostatecznym celem analiz lub krokiem w większym projekcie analiz.

Anomalie mają zwykle dwa źródła: rzeczywiste skrajne lub nietypowe wydarzenia oraz błędy z etapu zbierania lub przetwarzania danych. Wiele etapów wykrywania wartości odstających wygląda tak samo niezależnie od źródła problemu, jednak sposób radzenia sobie z konkretnymi anomaliami zależy od ich przyczyny. Dlatego poznanie powodu anomalii i rozróżnianie dwóch rodzajów ich źródeł jest ważne w procesie analiz.

Rzeczywiste zdarzenia mogą generować wartości odstające z różnych powodów. Anomalie w danych mogą wskazywać na oszustwo, włamanie do sieci, strukturalne usterki w produkcie, luki w strategiach lub zastosowanie produktu w sposób, który nie był oczekiwany ani przewidziany przez programistów. Wykrywanie anomalii często stosuje się do wykrywania oszustw finansowych. Także w dziedzinie cyberbezpieczeństwa przeprowadza się analizy tego rodzaju. Anomalie czasem pojawiają się nie dlatego, że napastnik próbuje zaatakować system, lecz za sprawą użytkownika korzystającego z produktu w nieoczekiwany sposób. Znam na przykład osobę, która używała aplikacji monitorującej aktywność fizyczną (bieganie, jazdę na rowerze, spacery itp.) do rejestrowania danych z jazd na torze do wyścigów samochodowych. Znajomy nie znalazł lepszej aplikacji i nie zastanawiał się nad tym, jak dziwnie szybkość samochodu na torze i pokonana tak odległość wyglądają w zestawieniu z wynikami zarejestrowanymi w trakcie jazdy na rowerze lub biegania. Gdy źródłem anomalii są rzeczywiste procesy, decyzja o tym, co z zrobić z danymi, wymaga solidnego zrozumienia przeprowadzanych analiz, wiedzy z dziedziny, znajomości zasad użytkowania, a czasem także obowiązującego systemu prawnego.

Anomalie mogą się też pojawić z powodu błędów w zbieraniu lub przetwarzaniu danych. W ręcznie wprowadzanych danych często występują literówki lub nieprawidłowe dane. Zmiany w formularzach, polach lub regułach sprawdzania poprawności mogą skutkować nieoczekiwanymi wartościami, w tym wartościami `null`. Firmy często monitorują działanie aplikacji internetowych i mobilnych, jednak wszelkie zmiany w tym, jak i gdzie rejestrowanie zdarzeń się odbywa, mogą skutkować anomaliami. Spędziłam wystarczająco dużo godzin na diagnozowaniu zmian we wskaźnikach, aby się

nauczyć, że warto najpierw spytać, czy niedawno zostały wprowadzone jakieś zmiany w rejestrowaniu zdarzeń. Przetwarzanie danych może skutkować powstawaniem wartości odstających na przykład wtedy, gdy jakieś wartości zostają błędnie odfiltrowane, etap przetwarzania nie zostanie zakończony lub dane zostaną wczytane kilkakrotnie, przez co powstaną duplikaty. Jeśli anomalie są wynikiem przetwarzania danych, zwykle łatwiej jest poprawić lub wyeliminować błędne wartości. Oczywiście zawsze, gdy jest to możliwe, warto usprawnić wcześniejsze etapy związane z wprowadzaniem lub przetwarzaniem danych, aby zapobiec późniejszym problemom z jakością.

W tym rozdziale najpierw omawiam kilka powodów do stosowania SQL-a w analizach tego rodzaju. Wyjaśniam też, kiedy SQL się nie sprawdza. Dalej przedstawiam zbiór danych o trzęsieniach ziemi, który będzie używany w przykładach w tym rozdziale. Opisuję także podstawowe narzędzia SQL-owe służące do wykrywania wartości odstających. Następnie omawiam różne rodzaje wartości odstających, do których wykrywania można stosować poszczególne narzędzia. Po wykryciu i zrozumieniu anomalii następny etap polega na ustaleniu, co z nimi zrobić. W obszarach wykrywania oszustw i cyberataków lub w monitorowaniu stanu zdrowia anomalie oznaczają problem, jednak w innych dziedzinach nie zawsze tak jest. Techniki z tego rozdziału można też stosować do wykrywania wyjątkowo cennych klientów, skutecznych kampanii marketingowych lub pozytywnych zmian w zachowaniach klientów. Czasami celem wykrywania anomalii jest ich przekazanie innym osobom lub maszynom, które mają rozwiązać problem. Jednak nierzadko znajdowanie wartości odstających to etap szerszych analiz, dlatego rozdział kończę przeglądem różnych sposobów korygowania anomalii.

Możliwości i ograniczenia SQL-a w zakresie wykrywania anomalii

SQL jest wszechstronnym językiem pozwalającym wykonywać różne zadania związane z analizą danych, jednak nie wszystko można w nim zrobić. W obszarze wykrywania anomalii SQL ma wiele zalet, ale też kilka wad, które sprawiają, że inne języki lub narzędzia mogą lepiej się nadawać do wykonywania niektórych operacji.

Zastosowanie SQL-a warto rozważyć, gdy dane już znajdują się w bazie, tak jak wcześniej w przypadku analizy szeregów czasowych i tekstu w rozdziałach 3. i 5. SQL wykorzystuje możliwości obliczeniowe bazy do szybkiego wykonywania obliczeń na wielu rekordach. Zwłaszcza gdy tabele z danymi są duże, przesyłanie danych z bazy do innego narzędzia jest czasochłonne. Praca w ramach bazy jest jeszcze bardziej uzasadniona, kiedy wykrywanie anomalii jest etapem większych analiz przeprowadzanych w SQL-u. Kod napisany w SQL-u można sprawdzić, aby zrozumieć, dlaczego określone rekordy zostały oznaczone jako wartości odstające. Ponadto kod w SQL-u pozostaje taki sam, nawet jeśli dane przesyłane do bazy się zmienią.

Jeżeli chodzi o wady, SQL nie ma zaawansowanych statystycznych funkcji dostępnych w pakietach opracowanych dla języków takich jak R i Python. Udostępnia kilka standardowych funkcji statystycznych, jednak dodatkowe, bardziej skomplikowane obliczenia statystyczne mogą być zbyt powolne lub wymagające dla niektórych baz. W sytuacjach wymagających bardzo szybkich odpowiedzi, na przykład przy wykrywaniu oszustw lub włamań, analizowanie danych w bazie może być nieakceptowalne, ponieważ dane są wczytywane z opóźnieniem (dotyczy to przede wszystkim analitycznych baz danych).

Często stosowany proces polega na użyciu SQL-a do wykonywania wstępnych analiz i określania wartości minimalnych, maksymalnych i średnich. Następnie opracowywany jest system monitorowania działający w czasie bardziej zbliżonym do rzeczywistego. W takich systemach wykorzystuje się usługi strumieniowania danych lub specjalne magazyny danych działające w czasie rzeczywistym. Można jednak badać w bazie wzorce anomalii, a następnie wykrywać je w usługach strumieniowania danych lub magazynach danych działających w czasie rzeczywistym. Następna wada związana jest z tym, że kod w SQL-u jest oparty na regułach, co opisałam w rozdziale 5. Bardzo dobrze nadaje się do analizowania znanego zestawu warunków lub kryteriów, ale nie dostosowuje się automatycznie do zmieniających się wzorców odpowiadających coraz to nowym napastnikom. W takich zastosowaniach często lepiej sprawdzają się techniki uczenia maszynowego i powiązane z nimi języki.

Po omówieniu zalet SQL-a i sytuacji, w których warto go stosować, opisuję zbiór danych używany w przykładach w tym rozdziale, a następnie przechodzę do samego kodu.

Zbiór danych

Dane do przykładów z tego rozdziału to zbiór rekordów dotyczący wszystkich trzęsień ziemi zarejestrowanych przez US Geological Survey (USGS) w latach 2010–2020. USGS udostępnia te dane w wielu formatach, w tym za pomocą kanałów informacyjnych przesyłanych w czasie rzeczywistym (*https://earthquake.usgs.gov/earthquakes/feed*).

Ten zbiór danych obejmuje około 1,5 miliona rekordów. Każdy rekord reprezentuje jedno trzęsienie ziemi i zawiera między innymi znacznik czasu, lokalizację, siłę, głębokość ogniska i źródło informacji. Przykładowe dane są pokazane na rysunku 6.1. Kompletny słownik danych (*https://oreil.ly/NjgCt*) znajdziesz w witrynie organizacji USGS.

*	time	latitude	longitude	depth	mag	net	place	type	status
1	2011-03-11 05:46:24	38.297	142.373	29	9.1	official	2011 Great Tohoku Earthquake, Japan	earthquake	reviewed
2	2010-02-27 06:34:11	-36.122	-72.898	22.9	8.8	official	offshore Bio-Bio, Chile	earthquake	reviewed
3	2012-04-11 08:38:36	2.327	93.063	20	8.6	official	off the west coast of northern Sumatra	earthquake	reviewed
4	2015-09-16 22:54:32	-31.5729	-71.6744	22.44	8.3	us	48km W of Illapel, Chile	earthquake	reviewed
5	2013-05-24 05:44:48	54.892	153.221	598.1	8.3	us	Sea of Okhotsk	earthquake	reviewed
6	2012-04-11 10:43:10	0.802	92.463	25.1	8.2	us	off the west coast of northern Sumatra	earthquake	reviewed
7	2017-09-08 04:49:19	15.0222	-93.8993	47.39	8.2	us	101km SSW of Tres Picos, Mexico	earthquake	reviewed
8	2014-04-01 23:46:47	-19.6097	-70.7691	25	8.2	us	94km NW of Iquique, Chile	earthquake	reviewed
9	2018-08-19 00:19:40	-18.1125	-178.153	600	8.2	us	286km NNE of Ndoi Island, Fiji	earthquake	reviewed
10	2019-05-26 07:41:15	-5.8119	-75.2697	122.57	8	us	78km SE of Lagunas, Peru	earthquake	reviewed
11	2013-02-06 01:12:25	-10.799	165.114	24	8	us	76km W of Lata, Solomon Islands	earthquake	reviewed
12	2011-03-11 06:15:40	36.281	141.111	42.6	7.9	us	near the east coast of Honshu, Japan	earthquake	reviewed
13	2017-01-22 04:30:22	-6.2464	155.1718	135	7.9	us	35km WNW of Panguna, Papua New Guinea	earthquake	reviewed
14	2018-01-23 09:31:40	56.0039	-149.1658	14.06	7.9	us	280km SE of Kodiak, Alaska	earthquake	reviewed
15	2016-12-17 10:51:10	-4.5049	153.5216	94.54	7.9	us	54km E of Taron, Papua New Guinea	earthquake	reviewed
16	2014-06-23 20:53:09	51.8486	178.7352	109	7.9	us	19km SE of Little Sitkin Island, Alaska	earthquake	reviewed
17	2018-09-06 15:49:18	-18.4743	179.3502	670.81	7.9	us	102km ESE of Suva, Fiji	earthquake	reviewed
18	2016-12-08 17:38:46	-10.6812	161.3273	40	7.8	us	69km WSW of Kirakira, Solomon Islands	earthquake	reviewed
19	2016-11-13 11:02:56	-42.7373	173.054	15.11	7.8	us	54km NNE of Amberley, New Zealand	earthquake	reviewed
20	2015-05-30 11:23:02	27.8386	140.4931	664	7.8	us	189km WNW of Chichi-shima, Japan	earthquake	reviewed
21	2020-07-22 06:12:44	55.0715	-158.596	28	7.8	us	99 km SSE of Perryville, Alaska	earthquake	reviewed
22	2010-04-06 22:15:01	2.383	97.048	31	7.8	us	northern Sumatra, Indonesia	earthquake	reviewed

Rysunek 6.1. Fragment danych z tabeli earthquakes

Powodem trzęsień ziemi są nagłe ruchy mas skalnych wzdłuż uskoków między płytami tektonicznymi pokrywającymi ziemię. W obszarach znajdujących się blisko krawędzi takich płyt występuje znacznie więcej trzęsień ziemi i są one znacznie silniejsze. Pacyficzny Pierścień Ognia to region wokół Oceanu Pacyficznego, w którym występuje wiele trzęsień ziemi. W analizach w tym rozdziale będzie się pojawiać wiele miejsc z tego obszaru, w tym Kalifornia, Alaska, Japonia i Indonezja.

Magnituda to miara siły trzęsienia ziemi w jego ognisku odpowiadająca natężeniu fal sejsmicznych. Magnituda jest mierzona na skali logarytmicznej. To oznacza, że trzęsienie ziemi o magnitudzie 5 jest 10-krotnie silniejsze od trzęsienia o magnitudzie 4. Same pomiary trzęsień ziemi to fascynujący temat, jednak wykracza poza zakres tej książki. Witryna organizacji USGS (*https://earthquake.usgs.gov*) to dobry punkt wyjścia, jeśli chcesz dowiedzieć się więcej na ten temat.

Wykrywanie wartości odstających

Choć pojęcie anomalii (wartości odstających), czyli punktu danych znacznie odbiegającego od pozostałych, wydaje się proste, znajdowanie takich wartości w zbiorach danych bywa trudne. Pierwszy problem związany jest z ustaleniem, kiedy wartości lub punkty danych są typowe, a kiedy rzadkie. Druga trudność dotyczy wyznaczania wartości progowej określającej, czy dany element należy uznać za anomalię. Dla danych z tabeli `earthquakes` przeprowadzę profilowanie poziomów głębokości i magnitudy, aby lepiej zrozumieć, które wartości są typowe, a które wyjątkowe.

Zwykle im większy i bardziej kompletny jest zbiór danych, tym łatwiej jest ocenić, jakie dane rzeczywiście są anomaliami. Czasem dostępne są etykiety („prawdy podstawowe"), z których można skorzystać. Etykieta przeważnie zapisana jest w kolumnie zbioru danych, która określa, czy rekord jest standardowy, czy stanowi wartość odstającą. „Prawdy podstawowe" można uzyskać ze źródeł branżowych lub naukowych, a także z wcześniejszych analiz. Na tej podstawie można na przykład stwierdzić, że każde trzęsienie ziemi o magnitudzie powyżej 7 jest anomalią. W innych sytuacjach konieczne jest samodzielne zbadanie danych i ich ocena. W tym rozdziale przyjmuję, że analizowany zbiór danych jest wystarczająco duży, aby można było samodzielnie wyznaczyć poziom wartości odstających (choć oczywiście istnieją zewnętrzne źródła, w których można sprawdzić magnitudy typowych i ekstremalnych trzęsień ziemi).

Narzędzia do wykrywania wartości odstających na podstawie samych danych dzielą się na kilka kategorii. Po pierwsze, można sortować (porządkować) wartości występujące w danych. Tę technikę można opcjonalnie połączyć z klauzulami *`GROUP BY`*, aby znajdować wartości odstające na podstawie częstości wystąpień. Po drugie, można zastosować funkcje statystyczne SQL-a, aby znaleźć skrajne wartości po obu krańcach przedziału wartości. Po trzecie, można wyświetlić dane na wykresie i zbadać je wizualnie.

Wyszukiwanie anomalii za pomocą sortowania

Jednym z podstawowych narzędzi do znajdowania wartości odstających jest sortowanie danych. Wykonuje się je za pomocą klauzuli *`ORDER BY`*. Ta klauzula domyślnie sortuje dane rosnąco (*`ASC`*). Aby posortować dane malejąco, użyj modyfikatora *`DESC`* po nazwie kolumny. Klauzula *`ORDER BY`* może dotyczyć jednej kolumny lub większej ich liczby. Każdą kolumnę można sortować rosnąco

lub malejąco niezależnie od innych. Sortowanie rozpoczyna się od pierwszej wskazanej kolumny. Jeśli podana jest druga kolumna, wyniki pierwszego sortowania są następnie porządkowane na podstawie drugiej kolumny (z zachowaniem pierwszego sortowania). Proces ten jest powtarzany dla wszystkich kolumn z klauzuli.

Ponieważ sortowanie ma miejsce po przetworzeniu przez bazę reszty zapytania, wiele baz umożliwia podawanie kolumn w zapytaniu nie tylko za pomocą nazw, ale też za pomocą numerów. Wyjątkiem jest baza SQL Server, gdzie trzeba stosować pełne nazwy. Ja preferuję składnię z numerami, ponieważ kod jest wtedy bardziej zwięzły, zwłaszcza gdy kolumny zapytania wymagają długich obliczeń lub wywołań funkcji.

Można na przykład posortować tabelę `earthquakes` na podstawie pola `mag` (zawiera ono wartość magnitudy):

```
SELECT mag
FROM earthquakes
ORDER BY 1 desc
;

mag
------
(null)
(null)
(null)
...
```

To powoduje zwrócenie wielu wierszy z wartością `null`. Warto zauważyć, że w tym zbiorze danych magnituda może mieć wartość `null`, którą też można uznać za wartość odstającą. Można usunąć wartości `null`:

```
SELECT mag
FROM earthquakes
WHERE mag is not null
ORDER BY 1 desc
;

mag
---
9.1
8.8
8.6
8.3
```

Występuje tylko jedna wartość powyżej 9 i dwie dodatkowe wartości większe niż 8,5. W wielu kontekstach te wartości nie wyglądają na duże. Jednak dzięki wiedzy na temat trzęsień ziemi można stwierdzić, że te poziomy są bardzo wysokie i nietypowe. Organizacja USGS utworzyła listę 20 największych trzęsień ziemi na świecie (*https://oreil.ly/gHUhy*). Wszystkie one mają magnitudę od 8,4 wzwyż, a tylko pięć ma magnitudę od 9,0 wzwyż. Trzy trzęsienia z tej listy miały miejsce w okresie od 2010 do 2020 roku (tego przedziału dotyczy analizowany zbiór danych).

Innym sposobem na ocenę, czy wartości są anomaliami w zbiorze danych, jest obliczanie częstości ich występowania. Można zliczyć wartości z pól `id` i pogrupować dane według magnitudy, aby ustalić, ile trzęsień ziemi miało określoną magnitudę. Liczbę trzęsień ziemi na magnitudę można następnie

podzielić przez łączną liczbę trzęsień obliczoną za pomocą funkcji okna `sum`. Wszystkie funkcje okna wymagają klauzuli *`OVER`* z klauzulą *`PARTITION BY`* i/lub *`ORDER BY`*. Ponieważ w mianowniku należy uwzględnić wszystkie rekordy, używam klauzuli *`PARTITION BY`* `1`, dzięki czemu funkcja okna wczytuje dane z całej tabeli. W ostatnim kroku zbiór wyników jest porządkowany na podstawie magnitudy:

```
SELECT mag
,count(id) as earthquakes
,round(count(id) * 100.0 / sum(count(id)) over (partition by 1),8)
 as pct_earthquakes
FROM earthquakes
WHERE mag is not null
GROUP BY 1
ORDER BY 1 desc
;

mag  earthquakes  pct_earthquakes
---  -----------  ---------------
9.1  1            0.00006719
8.8  1            0.00006719
8.6  1            0.00006719
8.3  2            0.00013439
...  ...          ...
6.9  53           0.00356124
6.8  45           0.00302370
6.7  60           0.00403160
...  ...          ...
```

Odnotowano tylko po jednym trzęsieniu ziemi o magnitudach powyżej 8,5, ale dwa o magnitudzie 8,3. Przy poziomie 6,9 liczba trzęsień ziemi jest dwucyfrowa, ale nadal reprezentują one bardzo niewielki procent danych. W analizach trzeba sprawdzić także drugą skrajność — najmniejsze wartości. W tym celu należy posortować dane rosnąco zamiast malejąco:

```
SELECT mag
,count(id) as earthquakes
,round(count(id) * 100.0 / sum(count(id)) over (partition by 1),8)
 as pct_earthquakes
FROM earthquakes
WHERE mag is not null
GROUP BY 1
ORDER BY 1
;

mag    earthquakes  pct_earthquakes
---    -----------  ---------------
-9.99  258          0.01733587
-9     29           0.00194861
-5     1            0.00006719
-2.6   2            0.00013439
...    ...          ...
```

Na drugim końcu listy wartości –9,99 i –9 występują częściej niż można oczekiwać. Choć nie można wyciągnąć logarytmu z zera ani liczb ujemnych, sam logarytm może być liczba ujemną, gdy argument jest większy niż 0, ale mniejszy niż 1. Na przykład log(0,5) wynosi około –0,301. Wartości –9,99 i –9 oznaczają trzęsienia ziemi o bardzo niewielkich magnitudach. Można się zastanawiać, czy tak niewielkie

wstrząsy są w ogóle wykrywalne. Jeśli wziąć pod uwagę dużą liczbę wystąpień takich wartości, podejrzewam, że reprezentują one nieznaną magnitudę, a nie bardzo słabe trzęsienia, dlatego można potraktować je jako anomalie.

Obok posortowania wszystkich danych przydatne może być ich pogrupowanie według jednego lub kilku pól z atrybutami, aby znaleźć anomalie w podzbiorach danych. Możesz na przykład sprawdzić najwyższe i najniższe magnitudy zarejestrowane w określonych lokalizacjach z pola `place`:

```
SELECT place, mag, count(*)
FROM earthquakes
WHERE mag is not null
 and place = 'Northern California'
GROUP BY 1,2
ORDER BY 1,2 desc
;

place                mag   count
-------------------  ----  -----
Northern California  5.61
Northern California  4.73  1
Northern California  4.51  1
...                  ...   ...
Northern California  -1.1  7
Northern California  -1.2  2
Northern California  -1.6  1
```

Północna Kalifornia (Northern California) to najczęściej występująca lokalizacja w zbiorze danych. Przejrzenie tylko wycinka danych dotyczących tego miejsca pokazuje, że najwyższe i najniższe wartości nie są tu tak skrajne jak w całym zbiorze danych. Trzęsienia ziemi o magnitudzie powyżej 5,0 nie są rzadkością na świecie, ale w Północnej Kalifornii stanowią wartości odstające.

Wyszukiwanie anomalii na podstawie percentyli i odchylenia standardowego

Sortowanie i opcjonalne grupowanie danych, a następnie wizualne przeglądanie wyników to przydatna technika wykrywania anomalii, przede wszystkim wtedy, gdy w danych występują skrajne wartości. Jednak bez wiedzy z danej dziedziny może nie być oczywiste, że trzęsienia ziemi o magnitudzie 9,0 są anomalią. Ilościowa ocena skrajności punktów danych zwiększa ścisłość analiz. Można ją przeprowadzać na dwa sposoby: za pomocą percentyli lub odchylenia standardowego.

Percentyle reprezentują odsetek wartości w rozkładzie, które są mniejsze od danej wartości. Mediana rozkładu to wartość, względem której jedna połowa populacji ma mniejsze wartości, a druga połowa większe. Mediana jest tak często używana, że w wielu bazach SQL-owych (choć nie we wszystkich) dostępna jest funkcja `median` do jej wyznaczania. Można też obliczyć inne percentyle, na przykład percentyl 25% (oznacza punkt danych, od którego 25% wartości jest mniejszych, a 75% większych) lub 89% (określa punkt danych, od którego 89% wartości jest mniejszych, a 11% większych). Percentyle często stosuje się w tekstach akademickich, na przykład w ustandaryzowanych testach, ale można z nich korzystać w dowolnej dziedzinie.

SQL udostępnia funkcję okna `percent_rank`, która zwraca percentyl dla każdego wiersza grupy. Podobnie jak we wszystkich funkcjach okna porządek sortowania jest wyznaczany za pomocą klauzuli *`ORDER BY`*. Funkcja `percent_rank`, tak jak `rank`, nie przyjmuje żadnego argumentu. Działa na wszystkich wierszach zwracanych przez zapytanie. Oto jej podstawowa postać:

```
percent_rank() over (partition by ... order by ...)
```

Klauzule *PARTITION BY* i *ORDER BY* są opcjonalne, jednak funkcja wymaga podania jakiejś wartości w klauzuli *OVER*, a określenie porządku zawsze jest dobrym pomysłem. Aby znaleźć percentyl odpowiadający magnitudzie każdego trzęsienia ziemi w każdej lokalizacji, można najpierw w podzapytaniu obliczyć wartości funkcji `percent_rank` dla każdego wiersza, a następnie w zapytaniu zewnętrznym zliczyć wystąpienia każdej magnitudy. Warto zauważyć, że ważne jest, aby obliczyć wartości funkcji `percent_rank` najpierw, przed wykonaniem agregacji, aby w obliczeniach zostały uwzględnione powtarzające się wartości:

```
SELECT place, mag, percentile
,count(*)
FROM
(
    SELECT place, mag
    ,percent_rank() over (partition by place order by mag) as percentile
    FROM earthquakes
    WHERE mag is not null
    and place = 'Northern California'
) a
GROUP BY 1,2,3
ORDER BY 1,2 desc
;

place               mag   percentile            count
------------------  ----  --------------------  -----
Northern California  5.6   1.0                   1
Northern California  4.73  0.9999870597065141    1
Northern California  4.51  0.9999741194130283    1
...                  ...   ...                   ...
Northern California  -1.1  3.8820880457568775E-5  7
Northern California  -1.2  1.2940293485856258E-5  2
Northern California  -1.6  0.0                   1
```

W Północnej Kalifornii trzęsienia ziemi o magnitudzie 5,6 mają percentyl 100%, co oznacza, że wszystkie pozostałe wartości są mniejsze. Trzęsienia o magnitudzie –1,6 mają percentyl 0, co oznacza, że żadne inne punkty danych nie są mniejsze.

Oprócz obliczenia dokładnego percentyla dla każdego wiersza SQL może podzielić zbiór danych na określoną liczbę przedziałów i zwracać przedziały, do których należą poszczególne wiersze. Służy do tego funkcja `ntile`. Możesz na przykład podzielić zbiór danych na 100 przedziałów:

```
SELECT place, mag
,ntile(100) over (partition by place order by mag) as ntile
FROM earthquakes
WHERE mag is not null
and place = 'Central Alaska'
ORDER BY 1,2 desc
;

place           mag   ntile
--------------  ----  -----
Central Alaska  5.4   100
Central Alaska  5.3   100
Central Alaska  5.2   100
...             ...   ...
Central Alaska  1.5   79
```

```
...             ...   ...
Central Alaska  -0.5  1
Central Alaska  -0.5  1
Central Alaska  -0.5  1
```

W wynikach dotyczących Środkowej Alaski widać, że trzy trzęsienia ziemi o magnitudzie powyżej 5 należą do percentyla 100%, magnituda 1,5 odpowiada percentylowi 79%, a najmniejsze wartości, –0,5, odpowiadają percentylowi 1%. Po obliczeniu tych wartości można wyznaczyć poziomy graniczne każdego przedziału, używając funkcji `max` i `min`. W tym przykładzie tworzę cztery przedziały, aby można było wygodnie przedstawić dane, ale jako argument funkcji `ntile` można podać dowolną dodatnią liczbę całkowitą:

```
SELECT place, ntile
,max(mag) as maximum
,min(mag) as minimum
FROM
(
    SELECT place, mag
    ,ntile(4) over (partition by place order by mag) as ntile
    FROM earthquakes
    WHERE mag is not null
    and place = 'Central Alaska'
) a
GROUP BY 1,2
ORDER BY 1,2 desc
;

place           ntile  maximum  minimum
--------------  -----  -------  -------
Central Alaska  4      5.4      1.4
Central Alaska  3      1.4      1.1
Central Alaska  2      1.1      0.8
Central Alaska  1      0.8      -0.5
```

Najwyższy przedział, 4., który reprezentuje percentyle od 75% do 100%, ma największy zakres: od 1,4 do 5,4. Środkowe 50% wartości, czyli przedziały 2. i 3., mają niewielki zakres od 0,8 do 1,4.

Oprócz obliczania percentyli lub przedziałów dla każdego wiersza można wyznaczyć wartości odpowiadające określonym percentylom dla całego zbioru wyników zapytania. Służą do tego funkcje `percentile_cont` i `percentile_disc`. Są to funkcje okna, mają jednak składnię nieco odmienną od wcześniej omawianych funkcji okna, ponieważ wymagają klauzuli *`WITHIN GROUP`*. Oto specyfikacja tych funkcji:

```
percentile_cont(liczba) within group (order by nazwa_pola) over (partition by
nazwa_pola)
```

Argument *`liczba`* to wartość z przedziału od 0 do 1 reprezentująca zwracany percentyl. Na przykład 0,25 oznacza percentyl 25%. Klauzula *`ORDER BY`* określa pole, na podstawie którego funkcja ma zwrócić podany percentyl, a także sposób uporządkowania wartości. Podobnie jak we wszystkich klauzulach *`ORDER BY`* w SQL-u opcjonalnie można użyć modyfikatorów *`ASC`* i *`DESC`* (domyślnie używany jest ten pierwszy). Klauzula *`OVER (PARTITION BY…)`* jest opcjonalna i, co zaskakujące, niektóre bazy jej nie obsługują, dlatego jeśli natrafisz na błąd, zajrzyj do dokumentacji używanej bazy.

Funkcja `percentile_cont` zwraca interpolowaną (obliczoną) wartość, która odpowiada danemu percentylowi, ale może nie występować w bazie. Z kolei funkcja `percentile_disc` zwraca wartość ze zbioru danych najbliższą żądanemu percentylowi. W dużych zbiorach danych i w zbiorach ze stosunkowo ciągłym rozkładem wartości wyniki obu funkcji są zwykle zbliżone, warto jednak się zastanowić, która z nich będzie bardziej odpowiednia do przeprowadzanych analiz. Zobacz teraz, jak stosować te funkcje w praktyce. Poniższy kod oblicza magnitudy odpowiadające percentylom 25%, 50% (mediana) i 75% dla wszystkich magnitud różnych od `null` ze Środkowej Alaski:

```
SELECT
percentile_cont(0.25) within group (order by mag) as pct_25
,percentile_cont(0.5) within group (order by mag) as pct_50
,percentile_cont(0.75) within group (order by mag) as pct_75
FROM earthquakes
WHERE mag is not null
and place = 'Central Alaska'
;

pct_25  pct_50  pct_75
------  ------  ------
0.8     1.1     1.4
```

To zapytanie zwraca zestawienie żądanych percentyli dla zbioru danych. Warto zauważyć, że uzyskane liczby odpowiadają wartościom maksymalnym dla 1., 2. i 3. przedziału z poprzedniego przykładu. W tym samym zapytaniu można obliczyć percentyle dla różnych pól, zmieniając pole w klauzuli *`ORDER BY`*:

```
SELECT
percentile_cont(0.25) within group (order by mag) as pct_25_mag
,percentile_cont(0.25) within group (order by depth) as pct_25_depth
FROM earthquakes
WHERE mag is not null
and place = 'Central Alaska'
;

pct_25_mag  pct_25_depth
----------  ------------
0.8         7.1
```

Funkcje `percentile_cont` i `percentile_disc`, w odróżnieniu od innych funkcji okna, wymagają klauzuli *`GROUP BY`* na poziomie zapytania, gdy w zapytaniu występują dodatkowe pola. Załóżmy, że chcesz przeanalizować dwa obszary Alaski, co wymaga użycia pola `place`. Wtedy w zapytaniu trzeba podać to pole w klauzuli *`GROUP BY`*, a percentyle będą obliczane dla grup mających te same wartości tego pola:

```
SELECT place
,percentile_cont(0.25) within group (order by mag) as pct_25_mag
,percentile_cont(0.25) within group (order by depth) as pct_25_depth
FROM earthquakes
WHERE mag is not null
and place in ('Central Alaska', 'Southern Alaska')
GROUP BY place
;

place            pct_25_mag  pct_25_depth
---------------  ----------  ------------
Central Alaska   0.8         7.1
Southern Alaska  1.2         10.1
```

Za pomocą tych funkcji można znaleźć dowolne percentyle potrzebne w analizach. Ponieważ mediana jest tak często obliczana, w wielu bazach dostępna jest funkcja `median` przyjmująca tylko jeden argument — pole, dla którego ma zostać wyznaczona mediana. Jest to wygodna i znacznie prostsza składnia, jeśli jednak funkcja `median` jest niedostępna, ten sam wynik można uzyskać za pomocą funkcji `percentile_cont`.

Dla dużych zbiorów danych funkcje `percentile` i `median` mogą być powolne i wymagające obliczeniowo. Wynika to z tego, że baza musi posortować wszystkie rekordy i przypisać im pozycje (zwykle w pamięci). Producenci niektórych baz udostępniają wersje tych funkcji dające przybliżone wyniki (na przykład `approximate_percentile`), które działają znacznie szybciej i zwracają wyniki bardzo zbliżone do funkcji wykonujących obliczenia na całym zbiorze danych.

Znajdowanie percentyli i przedziałów dla wartości ze zbioru danych umożliwia bardziej ilościową ocenę anomalii. Dalej w rozdziale zobaczysz, że uzyskane w ten sposób wartości pomagają też radzić sobie z anomaliami w zbiorze danych. Jednak ponieważ percentyle zawsze są podawane na skali od 0 do 100, nie pozwalają ocenić, jak bardzo nietypowe są określone wartości. Aby to stwierdzić, można zastosować inne funkcje statystyczne dostępne w SQL-u.

Do pomiaru tego, jak skrajne są wartości w zbiorze danych, można posłużyć się *odchyleniem standardowym*. Jest ono miarą zmienności w zbiorze wartości. Niższa wartość oznacza mniejszą zmienność, a wyższa wartość — większą zmienność. Gdy dane mają rozkład normalny, około 68% wartości znajduje się w granicach od –1 do +1 odchylenia standardowego od średniej, a 95% znajduje się w granicach od –2 do +2 odchylenia standardowego od średniej. Odchylenie standardowe jest równe pierwiastkowi kwadratowemu z sumy różnic wartości od zmiennej podzielonej przez liczbę obserwacji:

$$\sqrt{\sum(x_i - \mu)^2/N}$$

W tym wzorze x_i to obserwacja, μ to średnia z wszystkich obserwacji, $\sum$ oznacza, że wszystkie wartości należy zsumować, a N to liczba obserwacji. Więcej informacji o obliczaniu odchylenia standardowego znajdziesz w dowolnym podręczniku do statystyki lub w witrynie internetowej poświęconej statystyce[1].

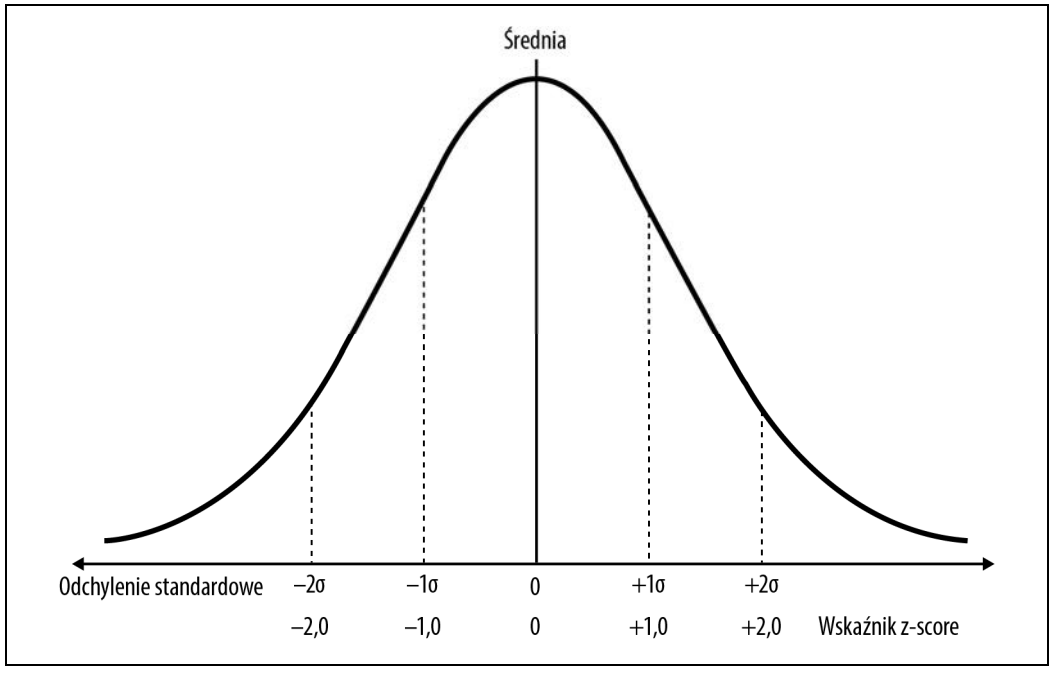
Większość baz udostępnia trzy funkcje do obliczania odchylenia standardowego. Funkcja `stddev_pop` wyznacza odchylenie standardowe dla populacji. Jeśli zbiór danych reprezentuje całą populację (jest to cecha wielu zbiorów z danymi o klientach), należy używać właśnie tej funkcji. Funkcja `stddev_samp` oblicza odchylenie standardowe dla próbki i różni się od pokazanego wzoru tym, że dzieli sumę przez $N - 1$, a nie przez N. Powoduje to zwiększenie odchylenia standardowego, co odzwierciedla utratę precyzji, gdy uwzględniana jest tylko część całej populacji. Dostępna w wielu bazach funkcja `stddev` działa identycznie jak `stddev_samp` i korzysta się z niej dlatego, że ma krótsza nazwę. Jeśli pracujesz ze zbiorem danych, który reprezentuje tylko próbkę, na przykład część ankiet lub wyników badań nad większą populacją, używaj funkcji `stddev_samp` lub `stddev`. W praktyce jeżeli pracujesz z dużymi zbiorami danych, różnice między wynikami zwracanymi przez funkcje `stddev_pop` i `stddev_samp` są zwykle niewielkie. Na przykład dla zawierającej 1,5 miliona rekordów tabeli `earthquakes` wartości zwracane przez te funkcję różnią się dopiero szóstą cyfrą po przecinku:

[1] Dobre wyjaśnienie znajdziesz na stronie *https://www.mathsisfun.com/data/standard-deviation-formulas.html*.

```
SELECT stddev_pop(mag) as stddev_pop_mag
,stddev_samp(mag) as stddev_samp_mag
FROM earthquakes
;

stddev_pop_mag stddev_samp_mag
------------------- -------------------
1.273605805569390395 1.273606233458381515
```

Te różnice są na tyle niewielkie, że w większości praktycznych zastosowań nie ma znaczenia, której funkcji użyjesz do obliczania odchylenia standardowego.

Za pomocą opisanych funkcji można obliczyć odchylenie standardowe od średniej dla każdej wartości ze zbioru danych. Ta wartość to wskaźnik *z-score* i służy do standaryzowania danych. Wartości powyżej średniej mają dodatnią wartość wskaźnika z-score, a wartości mniejsze od średniej mają wartość ujemną tego wskaźnika. Na rysunku 6.2 pokazane jest, jak wskaźnik z-score i odchylenie standardowe są powiązane z rozkładem normalnym.



Rysunek 6.2. Odchylenie standardowe i wskaźniki z-score dla rozkładu normalnego

Aby obliczyć wskaźnik z dla trzęsień ziemi, należy najpierw w podzapytaniu obliczyć średnią i odchylenie standardowe dla całego zbioru danych. Następnie wystarczy złączyć wynik ze zbiorem danych, tworząc iloczyn kartezjański, tak aby średnia i odchylenie standardowe zostały złączone z każdym wierszem reprezentującym trzęsienie ziemi. W tym celu należy użyć wyrażenia `1 = 1`, ponieważ większość baz wymaga zdefiniowania jakiegoś warunku złączenia.

W zewnętrznym zapytaniu odejmij średnią magnitudę od magnitudy każdego trzęsienia ziemi, a następnie podziel wynik przez odchylenie standardowe:

```
SELECT a.place, a.mag
,b.avg_mag, b.std_dev
,(a.mag - b.avg_mag) / b.std_dev as z_score
FROM earthquakes a
JOIN
(
    SELECT avg(mag) as avg_mag
    ,stddev_pop(mag) as std_dev
    FROM earthquakes
    WHERE mag is not null
) b on 1 = 1
WHERE a.mag is not null
ORDER BY 2 desc
;
```

```
place                                mag  avg_mag  std_dev  z_score
-----------------------------------  ---  -------  -------  -------
2011 Great Tohoku Earthquake, Japan  9.1  1.6251   1.2736   5.8691
offshore Bio-Bio, Chile              8.8  1.6251   1.2736   5.6335
off the west coast of northern Sumatra 8.6 1.6251  1.2736   5.4765
...                                  ...  ...      ...      ...
Nevada                               -2.5 1.6251   1.2736   -3.2389
Nevada                               -2.6 1.6251   1.2736   -3.3174
Nevada                               -2.6 1.6251   1.2736   -3.3174
```

Największe trzęsienia ziemi mają wskaźnik z-score równy prawie 6, a dla najmniejszych trzęsień (z wyłączeniem trzęsień o magnitudzie –9 i –9,99, które prawdopodobnie są anomalią związaną z wprowadzaniem danych) ten wskaźnik wynosi blisko 3. Można podsumować te wyniki tak, że największe trzęsienia ziemi są bardziej skrajnymi wartościami odstającymi niż najmniejsze trzęsienia.

Tworzenie wykresów w celu znajdowania anomalii

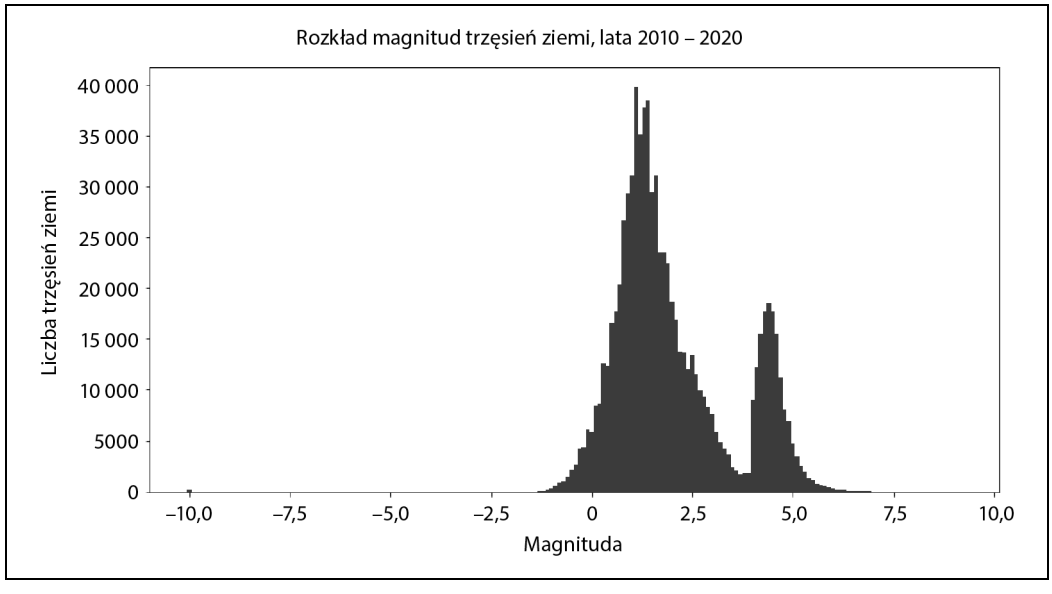
Obok znajdowania anomalii za pomocą sortowania danych oraz obliczania percentyli i odchyleń standardowych można też wizualizować dane na różnego rodzaju wykresach. W poprzednich rozdziałach pokazałam, że jedną z zalet wykresów jest możliwość podsumowywania i przedstawiania wielu punktów danych w zwięzłej postaci. Na podstawie analizy wykresów często można dostrzec wzorce i wartości odstające, które mogą nie być widoczne w surowych danych wyjściowych. Wykresy pomagają też w prezentowaniu danych i ewentualnych problemów związanych z anomaliami innym osobom.

W tym punkcie przedstawiam trzy rodzaje wykresów przydatne do wykrywania anomalii: wykresy słupkowe, wykresy punktowe i wykresy pudełkowe. Kod w SQL-u potrzebny do uzyskania takich wykresów jest prosty, choć może wymagać zastosowania opisanych w poprzednich rozdziałach technik tworzenia tabel przestawnych — zależy to od możliwości i ograniczeń oprogramowania używanego do tworzenia wykresów. Wszystkie popularne narzędzia do analizy biznesowej, arkusze kalkulacyjne oraz języki takie jak Python i R umożliwiają generowanie wykresów. Wykresy z tego protokołu zostały utworzone w Pythonie za pomocą biblioteki Matplotlib.

Wykres słupkowy służy do wyświetlania histogramu lub rozkładu wartości z pola i jest przydatny zarówno do zapoznawania się z charakterystyką danych, jak i do wykrywania wartości odstających. Jedna z osi reprezentuje cały zakres wartości, a na drugiej podawana jest liczba wystąpień każdej wartości. Interesujące są skrajnie niskie i skrajnie wysokie wartości, a także kształt wykresu. Można szybko ocenić, czy rozkład jest w przybliżeniu normalny (symetryczny wokół szczytowej lub średniej wartości), czy ma inny kształt albo kilka szczytów.

Aby wyświetlić histogram magnitud trzęsień ziemi, najpierw należy utworzyć zbiór danych pogrupowany według magnitud i z liczbą trzęsień przypisanych do poszczególnych grup. Następnie można wyświetlić dane wyjściowe pokazane na rysunku 6.3.

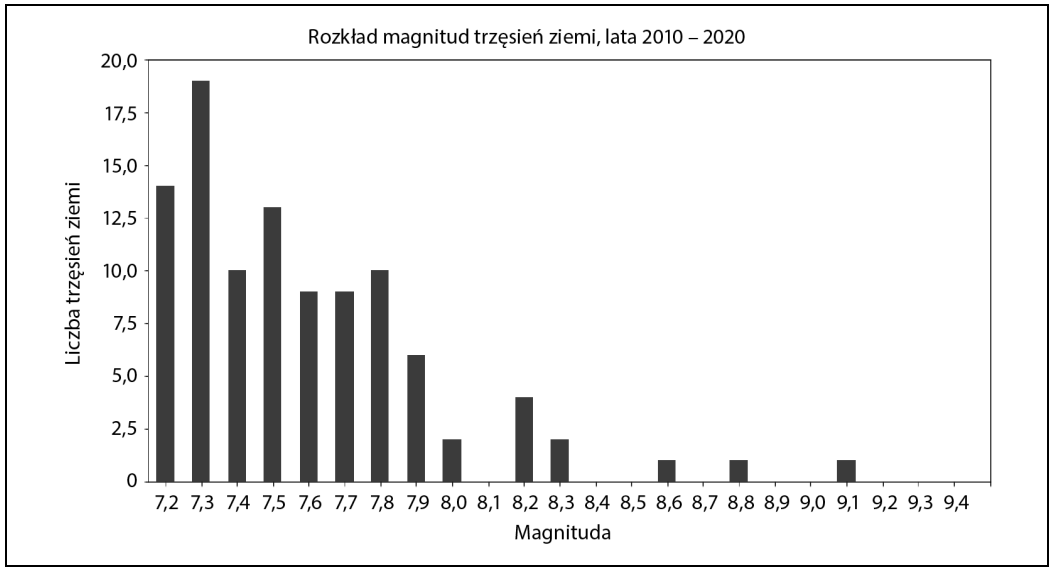
```
SELECT mag
,count(*) as earthquakes
FROM earthquakes
GROUP BY 1
ORDER BY 1
;

mag    earthquakes
-----  -----------
-9.99  258
-9     29
-5     1
...    ...
```



Rysunek 6.3. Rozkład magnitud trzęsień ziemi

Ten wykres obejmuje wartości od –10,0 do +10,0, co jest zgodne z wcześniejszą analizą danych. Na poziomie od 1,1 do 1,4 występuje szczyt z mniej więcej symetrycznym rozkładem po obu stronach, ale na poziomie 4,4 występuje drugi szczyt z mniej więcej 20 000 trzęsień ziemi. Przyczyny pojawienia się tego drugiego szczytu omawiam w następnym podrozdziale, który poświęcony jest rodzajom anomalii. Na tym wykresie trudno jest dostrzec skrajne wartości, dlatego warto przyjrzeć się bliżej fragmentowi wykresu (zobacz rysunek 6.4).

Tu łatwiej jest dostrzec częstość występowania najsilniejszych trzęsień ziemi, a także spadek częstości z ponad 10 (7 z małym ułamkiem) do 1 (powyżej 8). Na szczęście tak silne wstrząsy są niezwykle rzadkie.

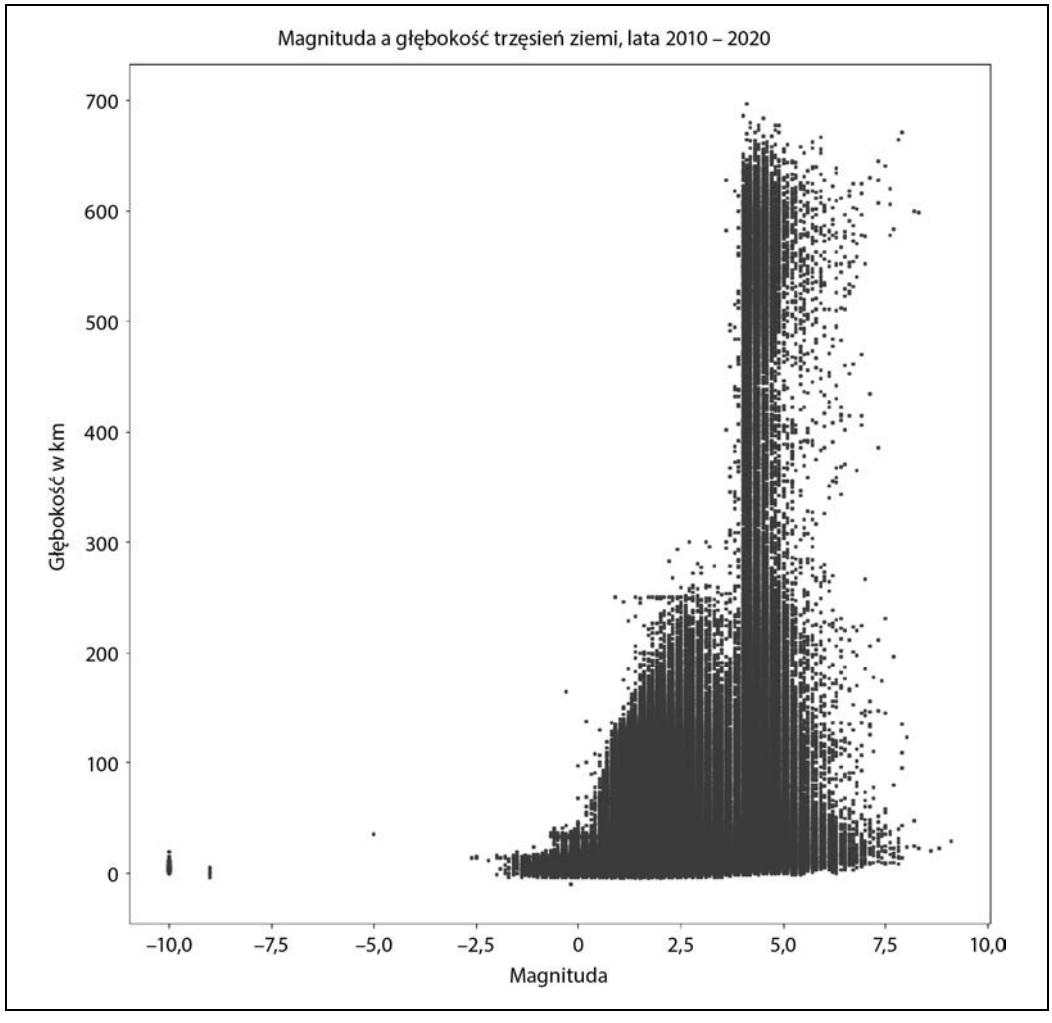


Rysunek 6.4. Przybliżenie rozkładu magnitud trzęsień ziemi obrazujące najwyższe magnitudy

Drugi rodzaj wykresu, który można wykorzystać do charakteryzowania danych i wykrywania wartości odstających, to *wykres punktowy*. Jest on przydatny, gdy zbiór danych obejmuje przynajmniej dwa istotne w analizach pola z wartościami liczbowymi. Na osi *x* wyświetlany jest zakres wartości z pierwszego pola, a na osi *y* — zakres wartości z drugiego pola. Na wykresie dla każdej pary wartości *x* i *y* ze zbioru danych rysowana jest kropka. Można na przykład wyświetlić magnitudę i głębokość trzęsień ziemi. Najpierw należy utworzyć zapytanie generujące zbiór danych z wszystkimi parami takich wartości. Następnie można wyświetlić uzyskane dane na wykresie (zobacz rysunek 6.5):

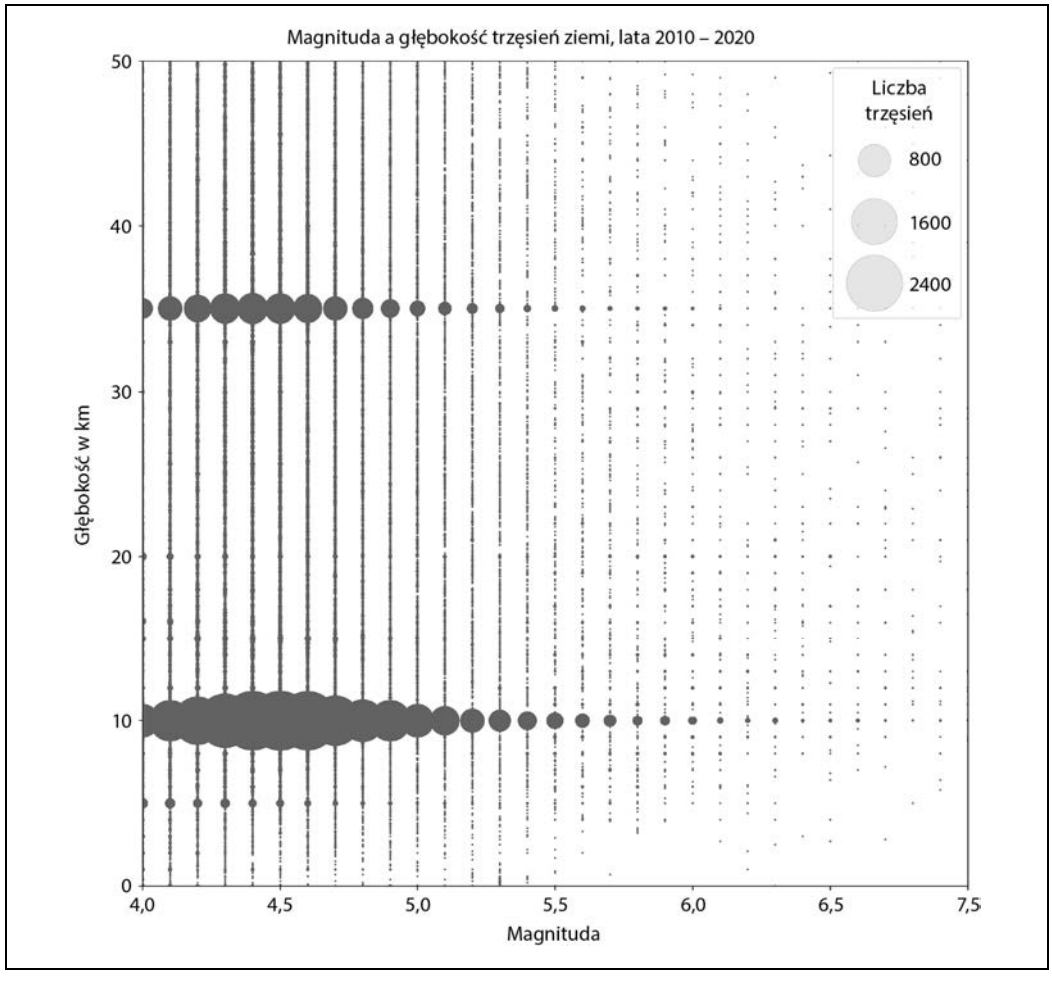
```
SELECT mag, depth
,count(*) as earthquakes
FROM earthquakes
GROUP BY 1,2
ORDER BY 1,2
;

mag    depth  earthquakes
-----  -----  -----------
-9.99  -0.59  1
-9.99  -0.35  1
-9.99  -0.11  1
...    ...    ...
```

Na tym wykresie widać ten sam zakres magnitud. Teraz jest on zestawiony z głębokościami, które wahają się od nieco poniżej 0 do około 700 km. Co ciekawe, trzęsienia o wysokiej głębokości (powyżej 300 km) odpowiadają magnitudom od mniej więcej 4 wzwyż. Niewykluczone, że tak głębokie trzęsienia ziemi można wykryć tylko wtedy, gdy przekraczają minimalną magnitudę. Warto zauważyć, że z powodu ilości danych zastosowałam uproszczenie i pogrupowałam wartości według kombinacji magnitudy oraz głębokości zamiast wyświetlać na wykresie całe 1,5 miliona punktów danych. Na podstawie liczby trzęsień ziemi można zmieniać wielkość poszczególnych punktów na wykresie, tak jak na rysunku 6.6, gdzie przybliżony jest zakres magnitud od 4,0 do 7,0 i zakres głębokości od 0 do 50 km.



Rysunek 6.5. Wykres punktowy z magnitudą i głębokością trzęsień ziemi

Trzecim rodzajem wykresów przydatnym w wyszukiwaniu i analizowaniu wartości odstających jest *wykres pudełkowy* (nazywany też *skrzynkowym* lub *ramkowym*). Na wykresach tego typu widoczne jest podsumowanie danych ze środkowego ich zakresu, przy czym zachowane zostają wartości odstające. Nazwa wykresu pochodzi od środkowego prostokąta przypominającego pudełko. Linia wyznaczająca podstawę prostokąta odpowiada percentylowi 25%, a linia wyznaczająca górę prostokąta znajduje się na poziomie percentyla 75%. Linia przechodząca przez środek odpowiada percentylowi 50% (medianie). Percentyle powinny być już zrozumiałe dzięki omówieniu z wcześniejszego podrozdziału. „Wąsy" wykresu pudełkowego to linie wychodzące z prostokąta. Ich wielkość to zwykle 1,5 rozstępu międzykwartylowego. *Rozstęp międzykwartylowy* to różnica między percentylami 75% i 25%. Wszystkie wartości wykraczające poza „wąsy" są wyświetlane na wykresie jako wartości odstające.



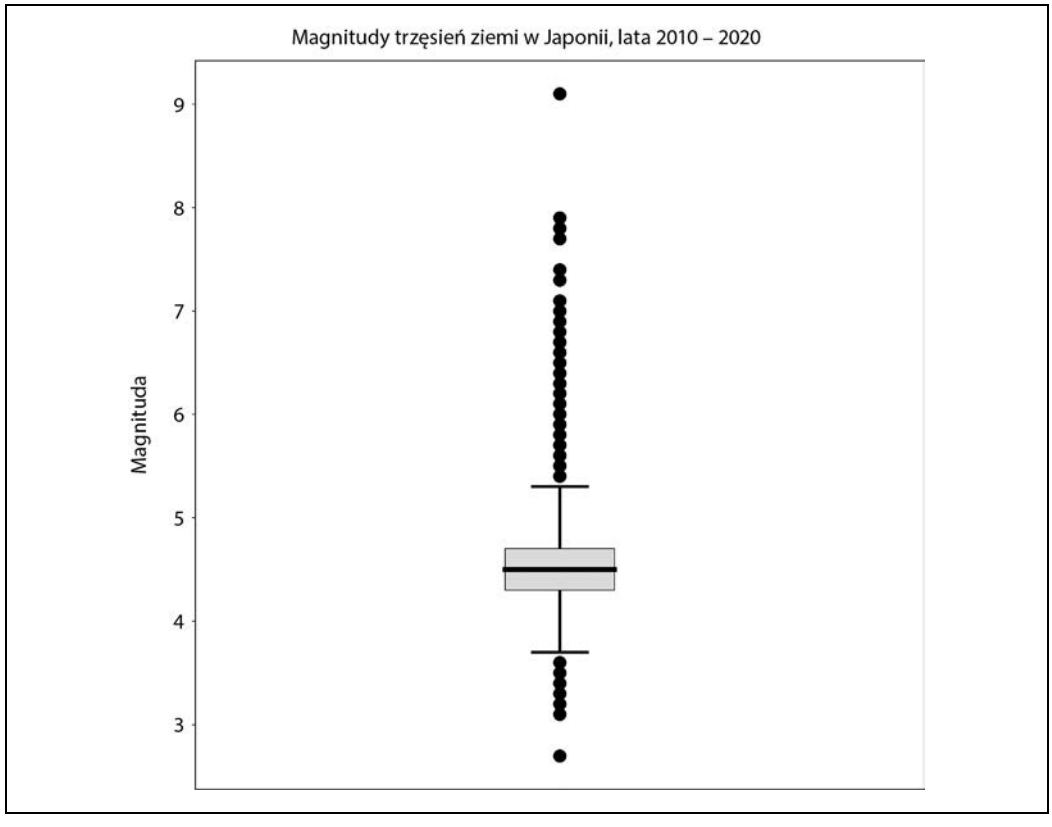
Rysunek 6.6. Wykres punktowy magnitudy i głębokości trzęsień ziemi; wersja przybliżona z wielkością punktów reprezentującą liczbę trzęsień

Niezależnie od tego, jakiego oprogramowania lub języka programowania używasz do wyświetlania wykresów pudełkowych, narzędzie automatycznie oblicza percentyle i rozstęp międzykwartylowy. Wiele narzędzi umożliwia też wyświetlanie „wąsów" na podstawie odchylenia standardowego od średniej lub według szerszych percentyli (na przykład 10% i 90%). Obliczenia zawsze odbywają się symetrycznie względem punktu środkowego (brane jest na przykład jedno odchylenie standardowe powyżej i poniżej średniej), jednak długość górnego i dolnego „wąsa" może być różna, ponieważ zależy od danych.

Na wykresie pudełkowym zazwyczaj uwzględniane są wszystkie wartości. Jednak ponieważ analizowany zbiór danych jest tak duży, w tym przykładzie wyświetlam tylko wykres podzbioru 16 036 trzęsień ziemi, które w polu `place` zawierają tekst „Japan". Najpierw tworzę potrzebny zbiór danych za pomocą SQL-a, używając prostej instrukcji *`SELECT`*, która pobiera wszystkie wartości `mag` spełniające warunki z filtra:

```
SELECT mag
FROM earthquakes
WHERE place like '%Japan%'
ORDER BY 1
;

mag
---
2.7
3.1
3.2
...
```

Następnie można utworzyć wykres pudełkowy w wybranym oprogramowaniu (zobacz rysunek 6.7).



Rysunek 6.7. Wykres pudełkowy z rozkładem magnitudy trzęsień ziemi w Japonii

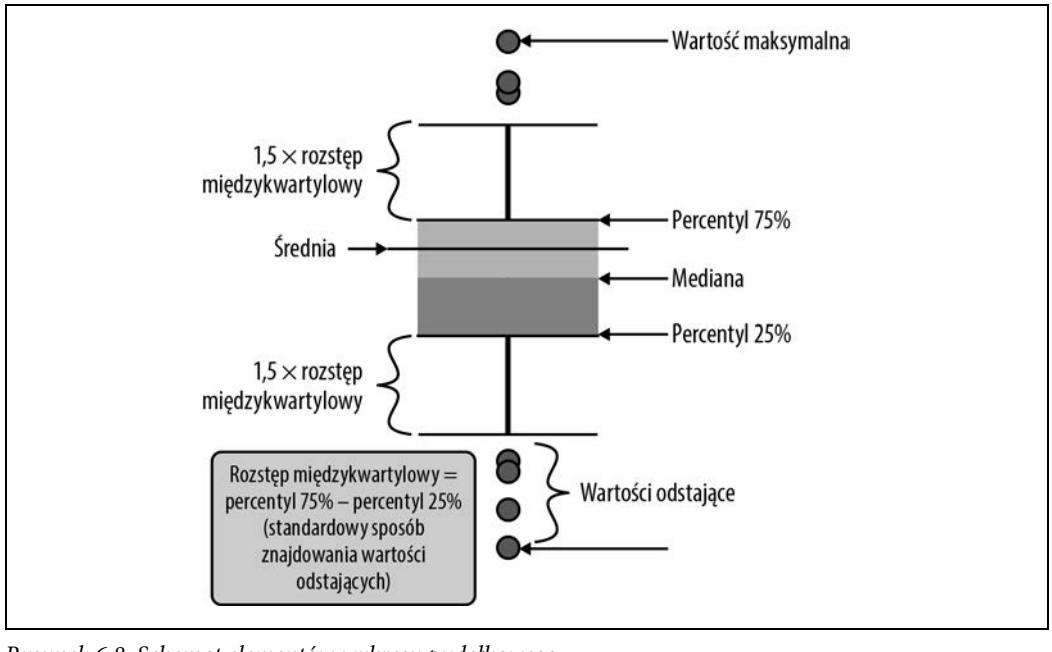
Choć potrzebne informacje często są obliczane za pomocą oprogramowania do tworzenia wykresów, najważniejsze wartości z wykresu pudełkowego można też uzyskać za pomocą SQL-a:

```
SELECT ntile_25, median, ntile_75
,(ntile_75 - ntile_25) * 1.5 as iqr
,ntile_25 - (ntile_75 - ntile_25) * 1.5 as lower_whisker
,ntile_75 + (ntile_75 - ntile_25) * 1.5 as upper_whisker
FROM
(
```

```
    SELECT
    percentile_cont(0.25) within group (order by mag) as ntile_25
    ,percentile_cont(0.5) within group (order by mag) as median
    ,percentile_cont(0.75) within group (order by mag) as ntile_75
    FROM earthquakes
    WHERE place like '%Japan%'
) a
;

ntile_25  median  ntile_75  iqr   lower_whisker  upper_whisker
--------  ------  --------  ----  -------------  -------------
4.3       4.5     4.7       0.60  3.70           5.30
```

Mediana magnitudy trzęsień ziemi w Japonii wynosi 4,5, a „wąsy" rozciągają się od 3,7 do 5,3. Punkty na wykresie reprezentują wartości odstające (zarówno bardzo słabe, jak i bardzo silne trzęsienia). Wielkie trzęsienie ziemi w Tohoku z 2011 roku miało magnitudę 9,1 i jest oczywistą anomalią — nawet wśród największych trzęsień ziemi w historii Japonii.

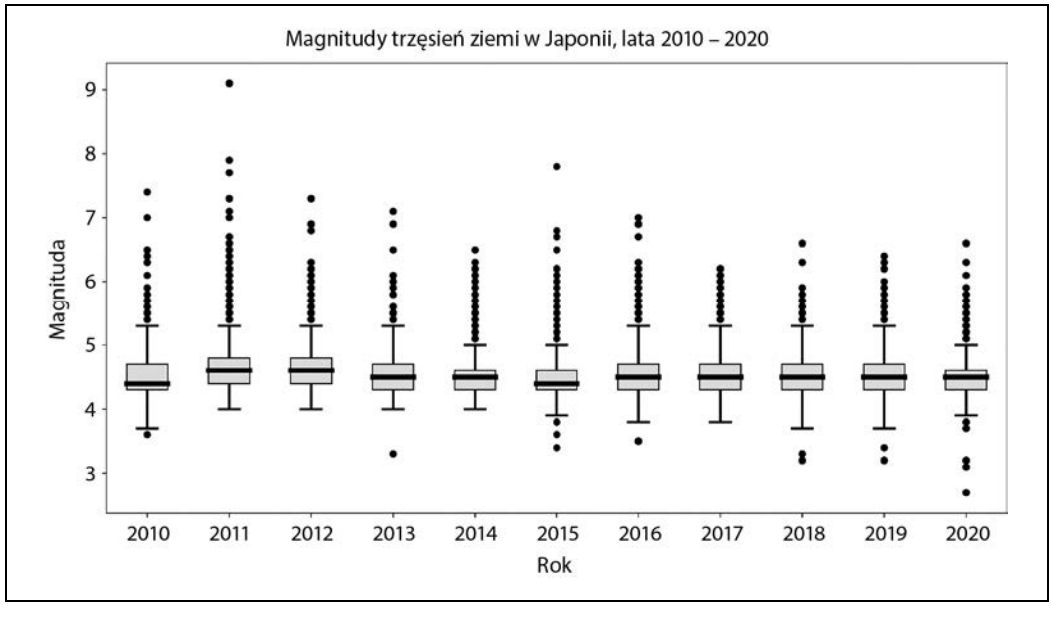
Z mojego doświadczenia wynika, że wykresy pudełkowe są jedną z wizualizacji, które najtrudniej jest wytłumaczyć osobom bez wiedzy statystycznej (i ludziom, którzy nie spędzają całych dni na przygotowywaniu i przeglądaniu wizualizacji). Trudny do zrozumienia jest przede wszystkim rozstęp międzykwartylowy, natomiast wartości odstające dla większości osób mają sens. Jeśli nie masz pewności, że odbiorcy będą potrafili zinterpretować wykres pudełkowy, postaraj się go wytłumaczyć w przejrzysty, nie nadmiernie techniczny sposób. Mam schemat podobny do tego z rysunku 6.8, na którym objaśniam elementy wykresu pudełkowego. Przesyłam ten schemat razem z wynikowym wykresem na wypadek, gdyby odbiorcy potrzebowali odświeżyć wiedzę.



Rysunek 6.8. Schemat elementów wykresu pudełkowego

Wykresy pudełkowe można też wykorzystać do porównywania grup danych, aby dodatkowo zidentyfikować i zdiagnozować miejsca występowania wartości odstających. Można na przykład porównać trzęsienia ziemi w Japonii w różnych latach. Najpierw należy dodać rok z pola `time` do danych wyjściowych w SQL-u, a następnie utworzyć wykres (zobacz rysunek 6.9):

```
SELECT date_part('year',time)::int as year
,mag
FROM earthquakes
WHERE place like '%Japan%'
ORDER BY 1,2
;

year  mag
----  ---
2010  3.6
2010  3.7
2010  3.7
...  ...
```



Rysunek 6.9. Wykres pudełkowy magnitud trzęsień ziemi w Japonii z podziałem na lata

Choć mediana i zakres zmieniają się nieco w poszczególnych latach, regularnie przyjmują wartości od 4 do 5. Co roku Japonię nawiedzają duże trzęsienia ziemi traktowane jako wartości odstające. W każdym roku było to przynajmniej jedno trzęsienie o magnitudzie powyżej 6,0, a w sześciu latach wystąpiło trzęsienie o magnitudzie 7,0 lub większej. Japonia jest więc bez wątpienia obszarem o bardzo dużej aktywności sejsmicznej.

Wykresy słupkowe, wykresy punktowe i wykresy pudełkowe są powszechnie używane do wykrywania i charakteryzowania wartości odstających w zbiorach danych. Umożliwiają szybkie przedstawienie złożonych aspektów dużych zbiorów danych i rozpoczęcie opowiadania historii na ich temat.

Razem z sortowaniem, percentylami i odchyleniem standardowym wykresy są ważnym elementem zestawu narzędzi do wykrywania anomalii. Po przedstawieniu tych narzędzi pora przejść do omówienia różnych rodzajów anomalii, które mogą się pojawiać obok opisanych do tego miejsca.

Rodzaje anomalii

Anomalie mogą przyjmować różne formy i rozmiary. W tym podrozdziale omawiam trzy ogólne kategorie anomalii: wartości, liczby wystąpień (częstość wystąpień) i obecność lub brak. Są to dobre punkty wyjścia do analizy każdego zbioru danych — czy to w ramach profilowania, czy to z powodu podejrzewania występowania anomalii. Wartości odstające i inne nieoczekiwane dane są często specyficzne dla dziedziny, dlatego zwykle im więcej wiadomo, jak i dlaczego dane są generowane, tym lepiej. Jednak opisane tu wzorce i techniki wykrywania anomalii są dobrym punktem wyjścia do dalszych analiz.

Anomalne wartości

Prawdopodobnie najczęściej występujący rodzaj anomalii i pierwsza rzecz, jaka przychodzi na myśl w tym kontekście, to pojedyncze wartości, które są skrajnie wysokimi lub niskimi wartościami odstającymi, albo nietypowe wartości w środkowej części rozkładu.

W poprzednim podrozdziale omówiłam kilka sposobów wyszukiwania wartości odstających: za pomocą sortowania, percentyli, odchylenia standardowego i wykresów. Okazało się, że w zbiorze danych o trzęsieniach ziemi występują zarówno niezwykle wysokie, jak i niezwykle niskie poziomy magnitudy. Magnitudy mają też różną liczbę *znaczących cyfr*, czyli cyfr po przecinku. Można na przykład sprawdzić podzbiór wartości zbliżonych do 1 i znaleźć wzorzec powtarzający się w analizowanym zbiorze danych:

```
SELECT mag, count(*)
FROM earthquakes
WHERE mag > 1
GROUP BY 1
ORDER BY 1
limit 100
;

mag         count
----------  -----
...         ...
1.08        3863
1.08000004  1
1.09        3712
1.1         39728
1.11        3674
1.12        3995
....        ...
```

Co jakiś czas pojawia się tu wartość z ośmioma znaczącymi cyframi. Wiele wartości ma dwie znaczące cyfry, ale częściej występuje tylko jedna znacząca cyfra. Prawdopodobnie wynika to z różnych poziomów precyzji instrumentów rejestrujących poziom magnitudy. Ponadto baza nie wyświetla drugiej znaczącej

cyfry, gdy jest nią zero. Na przykład wartość 1,10 jest wyświetlana jako 1,1. Jednak duża liczba rekordów o wartości 1,1 wskazuje na to, że nie chodzi tu tylko o sposób wyświetlania. W zależności od celu analiz można zaokrąglić wartości do tej samej liczby znaczących cyfr.

Oprócz wyszukiwania anomalnych wartości często pomocne jest też ustalenie, dlaczego wystąpiły, i wykrycie innych atrybutów skorelowanych z anomaliami. W tym obszarze potrzebne są kreatywność i umiejętności detektywistyczne. Na przykład w 1215 rekordach ze zbioru danych odnotowano bardzo dużą głębokość, przekraczającą 600 km. Warto ustalić, gdzie uzyskano te wartości odstające i w jaki sposób zostały zarejestrowane. Sprawdź źródło tych informacji. Można je znaleźć w polu `net` (od *network*, czyli sieć):

```
SELECT net, count(*)
FROM earthquakes
WHERE depth > 600
GROUP BY 1
;

net  count
---  -----
us   1215
```

Z witryny organizacji USGS można się dowiedzieć, że tym źródłem jest USGS National Earthquake Information Center, PDE (*https://earthquake.usgs.gov/data/comcat/contributor/us*). Nie jest to jednak zbyt pomocna informacja, dlatego warto sprawdzić wartości z pola `place`, które zawiera lokalizację trzęsień:

```
SELECT place, count(*)
FROM earthquakes
WHERE depth > 600
GROUP BY 1
;

place                          count
-----------------------------  -----
100km NW of Ndoi Island, Fiji  1
100km SSW of Ndoi Island, Fiji 1
100km SW of Ndoi Island, Fiji  1
...                            ...
```

Wizualna analiza wskazuje, że wiele bardzo głębokich trzęsień ziemi miało miejsce w okolicach wyspy Ndoi w archipelagu Fidżi. Jednak w lokalizacji podane są odległość i kierunek, na przykład „100km NW of", przez co podsumowanie tych danych jest utrudnione. Można przeprowadzić parsowanie tekstu, aby skupić się na danej lokalizacji, co pomoże w wyciągnięciu wniosków. Dlatego lokalizacje, które zawierają jakąś wartość, człon „ of ", a następnie dalsze wartości, należy podzielić na członie „ of " i pobrać drugą część:

```
SELECT
case when place like '% of %' then split_part(place,' of ',2)
     else place end as place_name
,count(*)
FROM earthquakes
WHERE depth > 600
GROUP BY 1
ORDER BY 2 desc
```

```
;

place_name         count
-----------------  -----
Ndoi Island, Fiji  487
Fiji region        186
Lambasa, Fiji      140
...                ...
```

Teraz można z większą pewnością stwierdzić, że większość bardzo głębokich trzęsień zarejestrowano w archipelagu Fidżi, a ich skupienie znajduje się w okolicy niewielkiej wyspy wulkanicznej Ndoi. Można kontynuować te analizy, na przykład parsując tekst, aby pogrupować wszystkie trzęsienia odnotowane w większym obszarze. Okaże się wtedy, że oprócz Fidżi inne bardzo głębokie trzęsienia zarejestrowano w okolicach wysp Vanuatu i Filipin.

Anomalie mogą mieć postać literówek, zapisów z różną wielkością liter i innych błędów tekstowych. Łatwość wykrywania takich sytuacji zależy od liczby różnych wartości w polu (czyli od *kardynalności* pola). Różnice w wielkości liter można wykryć, zliczając wszystkie unikatowe wartości i unikatowe wartości po zastosowaniu funkcji `lower` lub `upper`:

```
SELECT count(distinct type) as distinct_types
,count(distinct lower(type)) as distinct_lower
FROM earthquakes
;

distinct_types  distinct_lower
--------------  --------------
25              24
```

W polu `type` występują 24 różne wartości, ale 25 różnych zapisów. Aby znaleźć wszystkie wartości, można za pomocą obliczeń oznaczyć te wartości, których zapis małymi literami różni się od pierwotnych danych. Dodanie liczby rekordów dla każdego zapisu pomaga zrozumieć dane i zdecydować, co zrobić z różnymi wartościami:

```
SELECT type
,lower(type)
,type = lower(type) as flag
,count(*) as records
FROM earthquakes
GROUP BY 1,2,3
ORDER BY 2,4 desc
;

type       lower      flag   records
---------  ---------  -----  -------
...        ...        ...    ...
explosion  explosion  true   9887
ice quake  ice quake  true   10136
Ice Quake  ice quake  false  1
...        ...        ...    ...
```

Łatwo dostrzec anomalię w postaci wartości „Ice quake”, ponieważ jest to jedyna wartość, dla której obliczone pole `flag` jest równe `false`. Ponieważ istnieje tylko jeden rekord z takim zapisem, a 10 136 rekordów z zapisem małymi literami, można przyjąć, że nietypowy zapis należy pogrupować z pozostałymi

rekordami. Można też zastosować inne funkcje tekstowe, na przykład `trim` (jeśli podejrzewasz, że wartości zawierają dodatkowe spacje na początku lub na końcu) lub `replace` (gdy sądzisz, że niektóre wartości mogą być zapisane na różny sposób, na przykład jako cyfra „2” i słowo „dwa”).

Literówki są trudniejsze do wykrycia niż inne rozbieżności. Jeśli istnieje znany zestaw poprawnych wartości i zapisów, można go użyć do sprawdzenia danych, wykonując złączenie zewnętrzne z tabelą zawierającą prawidłowe wartości lub łącząc instrukcję `CASE` z listą operatora `IN`. W obu sytuacjach celem jest oznaczenie wartości, które są nieoczekiwane lub nieprawidłowe. Bez zestawu poprawnych wartości można albo wykorzystać wiedzę z dziedziny, albo posłużyć się domysłami. W tabeli `earthquakes` można sprawdzić wartości `type` występujące w bardzo nielicznych rekordach, a następnie spróbować ustalić, czy istnieje inna, częściej powtarzająca się wartość, którą należy zastosować zamiast pierwotnych danych:

```
SELECT type, count(*) as records
FROM earthquakes
GROUP BY 1
ORDER BY 2 desc
;

type                        records
--------------------------  -------
...                         ...
landslide                   15
mine collapse               12
experimental explosion      6
building collapse           5
...                         ...
meteorite                   1
accidental explosion        1
collapse                    1
induced or triggered event  1
Ice Quake                   1
rockslide                   1
```

Wcześniej sprawdziłam już wartość „Ice Quake” i uznałam, że prawdopodobnie chodzi o „ice quake”. Jest tylko jeden rekord z nazwą „rockslide”, choć można uznać, że jest ona zbliżona do innej wartości, „landslide”, która występuje w 15 rekordach. Trudniej ustalić znaczenie słowa „collapse”, ponieważ w zbiorze danych występują określenia „mine collapse” i „building collapse”. To, co należy z nimi zrobić (i czy w ogóle warto je modyfikować), zależy od celu analiz, co omawiam dalej, w punkcie „Radzenie sobie z anomaliami”.

Anomalne liczby wystąpień

Czasem anomalie mają postać nie tyle pojedynczych wartości, co wzorców lub klastrów aktywności w danych. Na przykład klient wydający w sklepie internetowym 100 złotych nie jest niczym wyjątkowym, ale jeśli ten sam klient wydaje 100 złotych co godzinę przez 48 godzin, oznacza to prawie na pewno anomalię.

Jest wiele wymiarów, na których klastry aktywności mogą wskazywać na anomalie (często zależy to od kontekstu danych). Czas i lokalizacja występują w wielu zbiorach danych, także w zbiorze `earthquakes`, dlatego posłużę się tymi wymiarami do zilustrowania technik omawianych w tym punkcie. Warto pamiętać, że te techniki często można stosować także do innych atrybutów.

Zdarzenia, które powtarzają się z nietypową częstością lub w krótkim czasie, mogą wskazywać na anomalną aktywność. Może to być korzystne — na przykład w sytuacji, gdy celebryta nieoczekiwanie zareklamuje jakiś produkt, co doprowadzi do gwałtownego wzrostu sprzedaży. Takie anomalie mogą też wskazywać na problem — na przykład nagły skok aktywności związany z używaniem fałszywej karty kredytowej lub próbą przeciążenia witryny. Aby zrozumieć tego rodzaju anomalie i móc ocenić, czy rzeczywiście stanowią one odstępstwo od zwykłych trendów, najpierw należy zastosować odpowiednie agregacje, a następnie posłużyć się technikami omówionymi wcześniej w tym rozdziale i metodami analizy szeregów czasowych przedstawionymi w rozdziale 3.

W następnych przykładach wykonuję serię kroków i zapytań, które pomagają zrozumieć typowe wzorce i wykryć niestandardowe sytuacje. Jest to iteracyjny proces, w którym w każdy kroku stosuję profilowanie danych, wiedzę z dziedziny i wnioski z wyników wcześniejszych zapytań. Analizy rozpoczynam od sprawdzenia liczby trzęsień w poszczególnych latach. W tym celu należy skrócić zawartość pola `time`, aby pozostawić sam rok, a następnie zliczyć rekordy. W bazie, w której funkcja `date_trunc` jest niedostępna, rozważ użycie funkcji `extract` lub `trunc`:

```
SELECT date_trunc('year',time)::date as earthquake_year
,count(*) as earthquakes
FROM earthquakes
GROUP BY 1
;

earthquake_year  earthquakes
---------------  -----------
2010-01-01       122322
2011-01-01       107397
2012-01-01       105693
2013-01-01       114368
2014-01-01       135247
2015-01-01       122914
2016-01-01       122420
2017-01-01       130622
2018-01-01       179304
2019-01-01       171116
2020-01-01       184523
```

Widać tu, że w latach 2011 i 2012 liczba trzęsień ziemi była niska w porównaniu z pozostałymi latami. Ponadto w 2018 roku nastąpił znaczny wzrost liczby trzęsień, który utrzymał się w latach 2019 i 2020. Wydaje się to dziwne. Można postawić hipotezę, że aktywność sejsmiczna ziemi nagle wzrosła, wystąpił błąd w danych (na przykład zduplikowanie rekordów) lub nastąpiły zmiany w procesie zbierania danych. Warto sprawdzić dane na poziomie miesięcznym, aby ustalić, czy ten sam trend jest widoczny także w krótszych przedziałach czasu:

```
SELECT date_trunc('month',time)::date as earthquake_month
,count(*) as earthquakes
FROM earthquakes
GROUP BY 1
;

earthquake_month  earthquakes
----------------  -----------
2010-01-01        9651
```

```
2010-02-01          7697
2010-03-01          7750
...                 ...
```

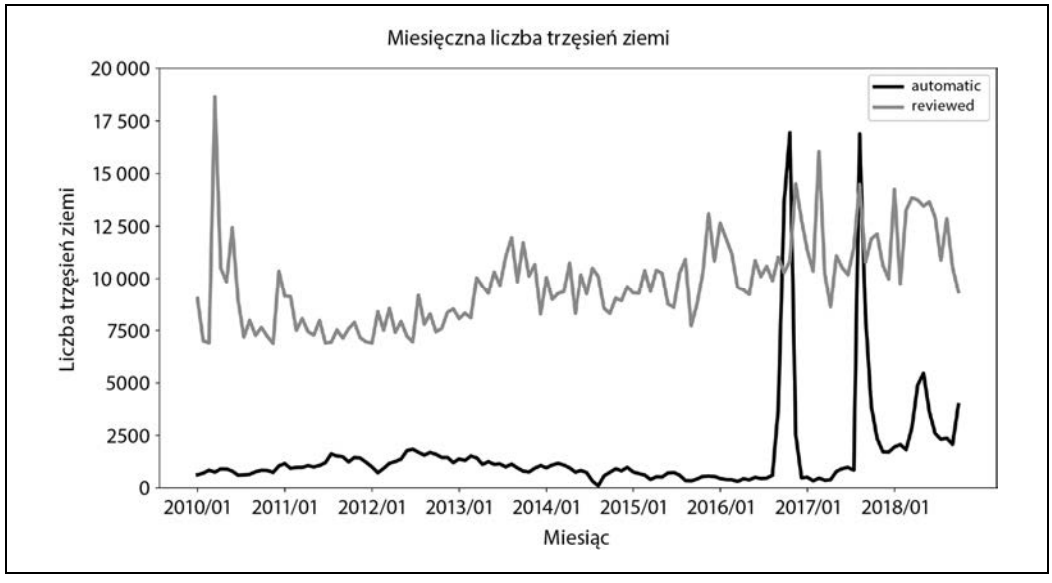
Dane wyjściowe są pokazane na rysunku 6.10. Widać tu, że choć liczba trzęsień ziemi zmienia się z miesiąca na miesiąc, to od 2017 roku następuje jej ogólny wzrost. Ponadto pojawiają się trzy odstające miesiące: kwiecień 2010 roku, lipiec 2018 roku i lipiec 2019 roku.

Rysunek 6.10. Liczba trzęsień ziemi w poszczególnych miesiącach

Następnie można sprawdzić dane w jeszcze krótszych przedziałach czasu i opcjonalnie przefiltrować zbiór danych na podstawie dat, aby skupić się na okresach, w których występują anomalie. Po ustaleniu dni, a nawet pór dnia, w których występują skoki liczby trzęsień ziemi, można przeanalizować dane na podstawie innych atrybutów. Może to pomóc w wyjaśnieniu anomalii, a przynajmniej w ustaleniu warunków, w jakich występują. Na przykład okazuje się, że wzrost liczby trzęsień ziemi widoczny od 2017 roku można przynajmniej częściowo wyjaśnić na podstawie pola `status`. Określa ono, czy zdarzenie zostało zweryfikowane przez człowieka (wartość „reviewed”), czy bezpośrednio dodane przez system bez oceny (wartość „automatic”):

```
SELECT date_trunc('month',time)::date as earthquake_month
,status
,count(*) as earthquakes
FROM earthquakes
GROUP BY 1,2
ORDER BY 1
;

earthquake_month  status     earthquakes
----------------  --------   -----------
2010-01-01        automatic  620
2010-01-01        reviewed   9031
2010-02-01        automatic  695
...               ...        ...
```

Trendy obrazujące liczbę wartości „automatic” i „reviewed” są pokazane na rysunku 6.11.



Rysunek 6.11. Liczba trzęsień ziemi miesięcznie z podziałem według statusu

Na wykresie widać, że odstająca liczba trzęsień ziemi w lipcu 2018 roku i lipcu 2019 roku wynika ze znacznego wzrostu raportów o statusie „automatic”, a w kwietniu 2010 roku wzrost wynikał z większej liczby zgłoszeń o statusie „reviewed”. Możliwe, że w 2017 roku do zbioru danych zaczęto dodawać zgłoszenia z automatycznych urządzeń pomiarowych. Możliwe też, że specjaliści nie mieli dość czasu na sprawdzenie wszystkich raportów.

Analizowanie w zbiorach danych lokalizacji, z których pochodzą informacje, to następny przydatny sposób na to, by wykryć i zrozumieć anomalie. Tabela `earthquakes` zawiera informacje o tysiącach bardzo małych trzęsień ziemi, które mogą utrudniać zrozumienie bardzo dużych i wartych uwagi wstrząsów. Przyjrzyj się teraz lokalizacji największych trzęsień ziemi, o magnitudzie co najmniej 6, i sprawdź, w jakich obszarach geograficznych się koncentrują:

```
SELECT place, count(*) as earthquakes
FROM earthquakes
WHERE mag >= 6
GROUP BY 1
ORDER BY 2 desc
;

place                                earthquakes
-----------------------------------  -----------
near the east coast of Honshu, Japan  52
off the east coast of Honshu, Japan   34
Vanuatu                               28
...                                   ...
```

Inaczej niż w przypadku czasu, gdzie badałam dane na coraz bardziej szczegółowym poziomie, wartości z pola `place` już są na tyle dokładne, że trudno jest zrozumieć ogólny obraz sytuacji, choć zdecydowanie wyróżnia się obszar „Honshu, Japan”. Można zastosować techniki analizy tekstu

z rozdziału 5., aby pobrać, a następnie pogrupować informacje geograficzne. Tu używam funkcji `split_part`, aby usunąć kierunek (na przykład „near the coast of" lub „100km N of"), który często pojawia się na początku pola `place`:

```
SELECT
case when place like '% of %' then split_part(place,' of ',2)
     else place
     end as place
,count(*) as earthquakes
FROM earthquakes
WHERE mag >= 6
GROUP BY 1
ORDER BY 2 desc
;

place                 earthquakes
--------------------  -----------
Honshu, Japan         89
Vanuatu               28
Lata, Solomon Islands 28
...                   ...
```

W regionie wokół wyspy Honsiu w Japonii („Honshu, Japan") miało miejsce 89 trzęsień ziemi, co oznacza, że jest to nie tylko lokalizacja największego trzęsienia w analizowanym zbiorze danych, ale też anomalia, jeśli chodzi o liczbę odnotowanych bardzo silnych wstrząsów. Możliwe jest też dalsze parsowanie, oczyszczanie i grupowanie wartości z pola `place` w celu uzyskania bardziej precyzyjnego obrazu występowania dużych trzęsień ziemi na świecie.

Proces wykrywania anomalnych liczb wystąpień, sum lub częstości w danych zwykle wymaga kilku serii zapytań na różnych poziomach szczegółowości. Często zaczyna się od ogólnego poziomu, następnie przechodzi do bardziej szczegółowego, potem znów na bardziej ogólny, aby porównać główne trendy, i ponownie do bardziej szczegółowego, by skupić się na określonych grupach lub wymiarach danych. Na szczęście SQL jest świetnym narzędziem do przeprowadzania tego rodzaju szybkich iteracji. Dodanie do tego technik badania szeregów czasowych (zobacz rozdział 3.) i tekstu (zobacz rozdział 5.) pozwala dodatkowo wzbogacić analizy.

Anomalie w postaci braku danych

Pokazałam już, że wyjątkowo duża liczba zdarzeń może wskazywać na anomalie. Warto jednak pamiętać, że także brak rekordów może sygnalizować odstępstwo od normy. Na przykład w trakcie operacji monitorowane jest tętno pacjenta. Brak tętna w jakimkolwiek momencie wywołuje alarm, podobnie jak nieregularności w pracy serca. Jednak w wielu kontekstach wykrywanie braku danych jest trudne (chyba że analityk celowo zwraca na tę kwestię uwagę). Klienci nie zawsze zgłaszają, że zamierzają zrezygnować z oferty firmy. Po prostu przestają korzystać z produktu lub usługi i niezauważalnie znikają ze zbioru danych.

Jednym ze sposobów na wykrywanie braku danych są techniki analizy kohortowej przedstawione w rozdziale 4. Złączenie szeregu lub wymiaru dat z danymi pozwala się upewnić, że dla każdej jednostki rekordy będą dostępne niezależnie od tego, czy dana jednostka była obecna w określonym przedziale czasu. Dzięki temu łatwiej jest wykryć nieobecność jednostek w danych.

Innym sposobem na wykrywanie braku danych są zapytania o luki, czyli o czas od ostatniego wystąpienia jednostki w danych. Niektóre obszary są bardziej narażone na silne trzęsienia ziemi z powodu układu płyt tektonicznych. Niektóre takie regiony wykryłam na podstawie danych we wcześniejszych przykładach. Trzęsienia ziemi trudno jest przewidzieć, nawet jeśli wiadomo, gdzie prawdopodobieństwo ich wystąpienia jest duże. Nie powstrzymuje to obserwatorów od spekulacji, gdzie może mieć miejsce następne wielkie trzęsienie. Często takie spekulacje są oparte na długim czasie, jaki upłynął od ostatniego trzęsienia. Można posłużyć się SQL-em, aby znaleźć odstępy między dużymi trzęsieniami ziemi i czas od ostatniego takiego wstrząsu:

```
SELECT place
,extract('days' from '2020-12-31 23:59:59' - latest)
 as days_since_latest
,count(*) as earthquakes
,extract('days' from avg(gap)) as avg_gap
,extract('days' from max(gap)) as max_gap
FROM
(
    SELECT place
    ,time
    ,lead(time) over (partition by place order by time) as next_time
    ,lead(time) over (partition by place order by time) - time as gap
    ,max(time) over (partition by place) as latest
    FROM
    (
        SELECT
        replace(
          initcap(
          case when place ~ ', [A-Z]' then split_part(place,', ',2)
               when place like '% of %' then split_part(place,' of ',2)
               else place end
        )
        ,'Region','')
        as place
        ,time
        FROM earthquakes
        WHERE mag > 5
    ) a
) a
GROUP BY 1,2
;

place             days_since_latest  earthquakes  avg_gap  max_gap
----------------  -----------------  -----------  -------  -------
Greece            62.0               109          36.0     256.0
Nevada            30.0               9            355.0    1234.0
Falkland Islands  2593.0             3            0.0      0.0
...               ...                ...          ...      ...
```

W wewnętrznym podzapytaniu kod parsuje i oczyszcza pole `place`, zwracając większe obszary lub państwa razem z czasem każdego trzęsienia ziemi. Uwzględniane są trzęsienia o magnitudzie co najmniej 5. W drugim podzapytaniu używam funkcji `lead` do znalezienia czasu następnego trzęsienia (jeśli takie miało miejsce) dla każdej kombinacji miejsca i czasu, a także do wyznaczenia przedziału czasu między każdą parą kolejnych trzęsień. Funkcja okna `max` zwraca ostatnie trzęsienie z poszczególnych miejsc. Zewnętrzne zapytanie oblicza dni od ostatniego trzęsienia o magnitudzie co najmniej 5 w zbiorze danych.

Używam do tego funkcji `extract`, aby zwracana była tylko liczba dni z przedziału tworzonego w wyniku odejmowania dwóch dat. Ponieważ w zbiorze danych występują tylko rekordy do końca 2020 roku, używam znacznika czasu „2020-12-31 23:59:59”. Gdy dane są na bieżąco odświeżane, lepiej jest użyć funkcji `current_timestamp` lub analogicznego wyrażenia. W podobny sposób liczba dni jest pobierana ze średniego i maksymalnego odstępu między trzęsieniami ziemi.

Czas od ostatniego dużego trzęsienia ziemi w danym miejscu może mieć w praktyce niewielką moc predykcyjną, jednak w wielu obszarach odstępy i czas od ostatniego wystąpienia jakiegoś zdarzenia mają praktyczne zastosowania. Obliczone typowe odstępy między aktywnościami wyznaczają poziom bazowy, z którym można porównać obecny czas od ostatniego zdarzenia. Jeśli znajduje się on w przedziale typowym dla wartości historycznych, można ocenić, że klient nie został utracony. Jeżeli jednak odstęp jest znacznie większy, ryzyko utraty klienta rośnie. Zbiór wyników zapytania zwracającego historyczne odstępy czasowe może się przydać do wykrywania anomalii, ponieważ pozwala na przykład ustalić najdłuższy czas, przez jaki klient nie korzystał z oferty firmy, ale potem wrócił.

Radzenie sobie z anomaliami

Anomalie mogą się pojawiać w zbiorach danych z różnych powodów i przyjmować wiele form. Po wykryciu anomalii w następnym kroku trzeba coś z nimi zrobić. Sposób traktowania anomalii zależy zarówno od ich źródła (problemów z procesem lub jakością danych), jak i od ostatecznego przeznaczenia zbioru danych lub analiz. Możliwości to: zbadanie anomalii bez wprowadzania zmian, wyeliminowanie ich, zastąpienie danych, przeskalowanie danych i wprowadzenie poprawek na wcześniejszych etapach prac.

Badanie anomalii

Ustalenie (lub próba znalezienia) przyczyn anomalii to zwykle pierwszy krok przy podejmowaniu decyzji, co z nimi zrobić. Ten aspekt procesu może być zarówno ciekawy, jak i frustrujący. Ciekawy w tym sensie, że badanie i rozwiązywanie zagadki wymaga umiejętności i kreatywności. Frustracja może wynikać z tego, że analitycy często pracują pod presją czasu, a badanie anomalii może przypominać drogę przez labirynt i skutkować wątpliwościami co do poprawności całego procesu analiz.

Gdy badam anomalie, zwykle wykonuję serię zapytań, w których na zmianę szukam wzorców lub przyglądam się konkretnym przykładom. Prawdziwe wartości odstające łatwo jest zauważyć. Wtedy przeważnie pobieram cały wiersz z wartością odstająca, aby poznać czas zdarzenia, źródło danych i inne dostępne atrybuty. Następnie sprawdzam rekordy o tych samych atrybutach, aby ustalić, czy też znajdują się w nich nietypowe wartości. Mogę na przykład sprawdzić, czy inne rekordy z tego samego dnia zawierają zwykłe czy niestandardowe dane. Użytkownicy kierowani do serwisu z określonej witryny lub transakcje dotyczące określonego produktu mogą pozwolić wykryć inne anomalie.

Jeśli dane są generowane w organizacji, w której pracuję, to po zbadaniu źródła i atrybutów anomalii kontaktuję się z interesariuszami lub właścicielami produktu. Czasem występuje znany błąd lub usterka, jednak często w procesie albo systemie istnieje poważny problem, którym trzeba się zająć. Przydatne są wtedy informacje o kontekście. W przypadku zewnętrznych lub publicznych zbiorów danych nie zawsze można znaleźć podstawową przyczynę problemu. Wtedy moim celem jest zebranie wystarczającej ilości informacji, aby zdecydować, które z opisanych dalej rozwiązań jest właściwe.

Usuwanie danych

Jednym ze sposobów na poradzenie sobie z anomaliami jest ich usunięcie ze zbioru danych. Jeśli są powody do podejrzeń, że w procesie zbierania danych wystąpił błąd wpływający na cały rekord, usunięcie danych jest odpowiednim podejściem. Jest to dobre rozwiązanie także wtedy, gdy zbiór danych jest na tyle duży, że usunięcie kilku rekordów prawdopodobnie nie wpłynie na wyciągnięte wnioski. Następnym dobrym powodem do usunięcia danych są skrajne wartości odstające, które mogłyby zaburzyć wyniki i doprowadzić do nieprawidłowych wniosków.

Wcześniej pokazałam, że zbiór danych o trzęsieniach ziemi zawiera sporo rekordów z magnitudą –9,99 i grupę rekordów z magnitudą –9. Ponieważ trzęsienia o tej sile byłyby niezwykle słabe, można podejrzewać, że są to błędne wartości lub dane wprowadzane w sytuacji, gdy magnituda jest nieznana. Można łatwo usunąć rekordy z takimi wartościami, używając warunku w klauzuli *WHERE*:

```
SELECT time, mag, type
FROM earthquakes
WHERE mag not in (-9,-9.99)
limit 100
;

time                 mag   type
-------------------  ----  ----------
2019-08-11 03:29:20  4.3   earthquake
2019-08-11 03:27:19  0.32  earthquake
2019-08-11 03:25:39  1.8   earthquake
```

Jednak przed usunięciem rekordów warto sprawdzić, czy uwzględnienie wartości odstających ma wpływ na dane wyjściowe. Możesz na przykład chcieć stwierdzić, czy usunięcie wartości odstających zmieni średnią magnitudę, ponieważ wartości odstające często zaburzają średnie. W tym celu należy obliczyć średnią dla całego zbioru danych, a także średnią po wykluczeniu skrajnie niskich wartości za pomocą instrukcji `CASE`:

```
SELECT avg(mag) as avg_mag
,avg(case when mag > -9 then mag end) as avg_mag_adjusted
FROM earthquakes
;

avg_mag             avg_mag_adjusted
------------------  ------------------
1.6251015161530643  1.6273225642983641
```

Średnie różnią się dopiero na trzeciej pozycji po przecinku (1,625 i 1,627), co oznacza niewielką różnicę. Jednak jeśli zastosować te same filtry tylko do lokalizacji Yellowstone National Park, w której występuje wiele wartości –9,99, różnica będzie bardziej widoczna:

```
SELECT avg(mag) as avg_mag
,avg(case when mag > -9 then mag end) as avg_mag_adjusted
FROM earthquakes
WHERE place = 'Yellowstone National Park, Wyoming'
;

avg_mag                 avg_mag_adjusted
----------------------  ----------------------
0.40639347873981053095  0.92332793709528214616
```

Choć wyniki nadal są niskie, różnica między średnimi 0,46 i 0,92 jest na tyle duża, że prawdopodobnie warto usunąć wartości odstające.

Dane można usuwać na dwa sposoby: albo w klauzuli *WHERE*, co powoduje wyeliminowanie wartości odstających z wszystkich wyników, albo w instrukcji CASE, co powoduje pominięcie ich tylko w określonych obliczeniach. To, które rozwiązanie należy wybrać, zależy od kontekstu analiz, a także od tego, czy ważne jest zachowanie wierszy na potrzeby zliczania wszystkich punktów danych lub wykorzystania przydatnych wartości z innych pól.

Zastępowanie innymi wartościami

Z anomalnymi wartościami często można radzić sobie przez zastępowanie ich innymi danymi (zamiast usuwania całych rekordów). Taką inną wartością może być domyślny zastępnik, najbliższa wartość liczbowa z określonego przedziału lub obliczona wartość statystyczna, na przykład średnia lub mediana.

Wcześniej pokazałam, że wartości null można zastępować wartością domyślną za pomocą funkcji coalesce. Gdy wartości są różne od null, ale z innych przyczyn powodują problemy, można je zastępować wartością domyślną w instrukcji CASE. Na przykład zamiast uwzględniać każdą aktywność sejsmiczną można pogrupować zdarzenia *niebędące* trzęsieniami ziemi w kategorii „Other”:

```
SELECT
case when type = 'earthquake' then type
     else 'Other'
     end as event_type
,count(*)
FROM earthquakes
GROUP BY 1
;

event_type  count
----------  -------
earthquake  1461750
Other       34176
```

To oczywiście powoduje zmniejszenie szczegółowości danych, ale może pomóc w podsumowaniu zbioru danych, w którym występuje wiele wartości odstających w polu type. Jeśli wiadomo, że wartości odstające są nieprawidłowe, zastępowanie ich w instrukcji CASE też pozwala zachować wszystkie wiersze w zbiorze danych. Możliwe na przykład, że na końcu rekordu dodano zbędne zero lub zarejestrowano wartość w calach zamiast w milach.

Inna technika radzenia sobie z liczbowymi wartościami odstającymi polega na zastępowaniu skrajnych wartości najbliższą wysoką lub niską liczbą, która nie jest odstająca. To podejście pozwala zachować dużą część przedziału wartości, ale zapobiega obliczaniu błędnych średnich, do jakich mogą prowadzić ekstremalne wartości odstające. *Winsoryzacja* to specjalna technika, która jest używana w tym celu. Polega ona na zastępowaniu wartości odstających wartościami odpowiadającymi określonemu percentylowi. Na przykład wartości powyżej percentyla 95% są zastępowane wartością odpowiadającą temu percentylowi i analogicznie modyfikowane są wartości poniżej percentyla 5%. Aby wykonać taką operację w SQL-u, należy najpierw obliczyć wartości percentyli 5% i 95%:

```
SELECT percentile_cont(0.95) within group (order by mag)
 as percentile_95
,percentile_cont(0.05) within group (order by mag)
```

```
 as percentile_05
FROM earthquakes
;

percentile_95  percentile_05
-------------  -------------
4.5            0.12
```

Można umieścić te obliczenia w podzapytaniu, a następnie użyć instrukcji `CASE` do zastępowania wartości odstających poniżej percentyla 5% i powyżej percentyla 95%. Zwróć uwagę na iloczyn kartezjański, który umożliwia zestawienie percentyli z poszczególnymi magnitudami:

```
SELECT a.time, a.place, a.mag
,case when a.mag > b.percentile_95 then b.percentile_95
      when a.mag < b.percentile_05 then b.percentile_05
      else a.mag
      end as mag_winsorized
FROM earthquakes a
JOIN
(
    SELECT percentile_cont(0.95) within group (order by mag)
     as percentile_95
    ,percentile_cont(0.05) within group (order by mag)
     as percentile_05
    FROM earthquakes
) b on 1 = 1
;

time                 place                        mag   mag_winsorize
-------------------  ---------------------------  ----  -------------
2014-01-19 06:31:50  5 km SW of Volcano, Hawaii   -9    0.12
2012-06-11 01:59:01  Nevada                       -2.6  0.12
...                  ...                          ...   ...
2020-01-27 21:59:01  31km WNW of Alamo, Nevada    2     2.0
2013-07-07 08:38:59  54km S of Fredonia, Arizona  3.5   3.5
...                  ...                          ...   ...
2013-09-25 16:42:43  46km SSE of Acari, Peru      7.1   4.5
2015-04-25 06:11:25  36km E of Khudi, Nepal       7.8   4.5
...                  ...                          ...   ...
```

Percentylowi 5% odpowiada wartość 0,12, a percentylowi 95% — 4,5. Wartości poniżej i powyżej tych poziomów progowych są zastępowane tymi poziomami, zapisywanymi przez kod w polu `mag_winsorize`. Wartości pomiędzy poziomami progowymi się nie zmieniają. W winsoryzacji nie ma określonych wartości progowych. Można użyć na przykład percentyli 1% i 99%, a nawet 0,01% i 99,9%. Zależy to od celu analiz, a także liczby wartości odstających i tego, jak bardzo są one ekstremalne.

Skalowanie

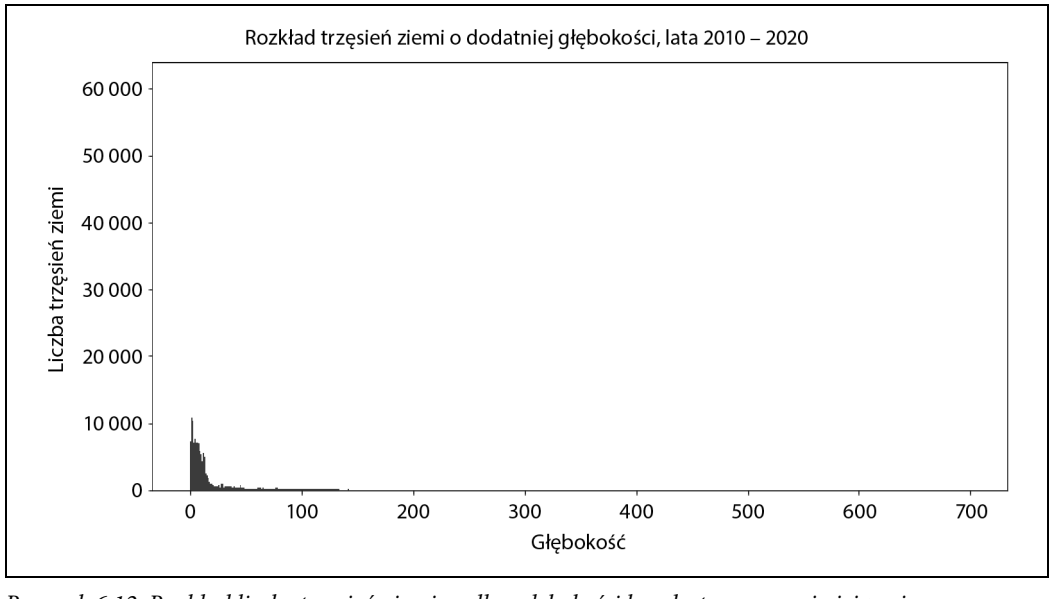
Zamiast odfiltrowywać rekordy lub modyfikować wartości odstające można przeskalować dane. Pozwala to zachować wszystkie wartości, a przy tym ułatwić analizy i tworzenie wykresów.

Wcześniej omówiłam już wskaźnik z-score. Tu warto wspomnieć, że można go zastosować do przeskalowania wartości. Ten wskaźnik jest przydatny, ponieważ można go używać zarówno do wartości dodatnich, jak i do wartości ujemnych.

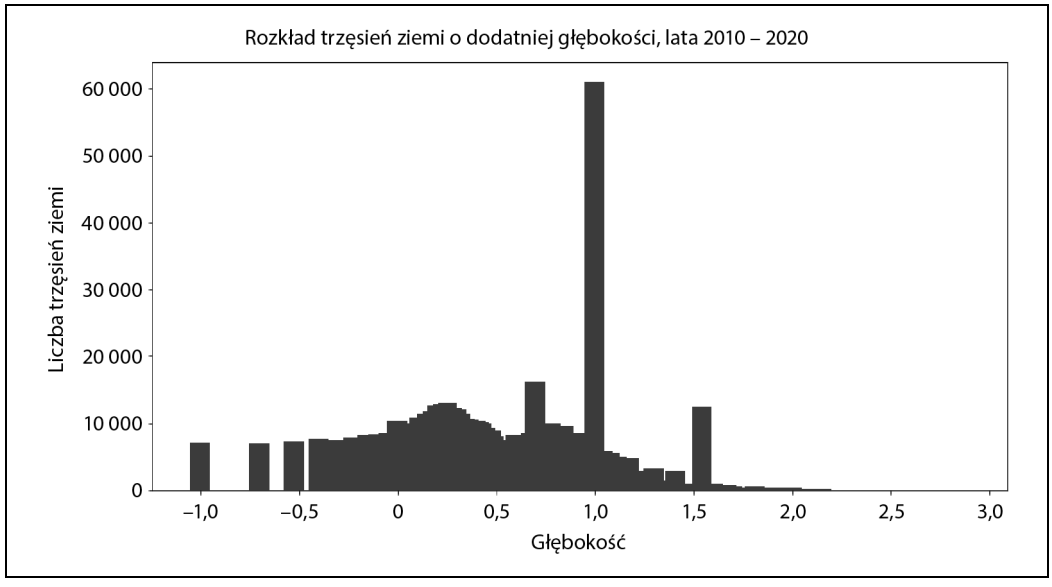
Inną często stosowaną transformacją jest przekształcanie wartości na skalę logarytmiczną. Zaletą tej techniki jest to, że zachowywana jest ta sama kolejność elementów, ale odległości między mniejszymi liczbami stają się relatywnie większe. Dane ze skali logarytmicznej można przekształcić z powrotem na pierwotną skalę, co ułatwia interpretację analiz. Wadą jest to, że tej metody nie można stosować do liczb ujemnych. W zbiorze danych o trzęsieniach ziemi magnituda jest już wyrażona na skali logarytmicznej. Magnituda 9,1 wielkiego trzęsienia ziemi w Tohoku jest bardzo wysoka, ale ta wartość odbiegałaby od normy jeszcze bardziej, gdyby nie była wyrażona na skali logarytmicznej.

Pole `deapth` zawiera wartości wyrażone w kilometrach. Następne zapytanie pobiera głębokość w pierwotnej postaci, a także po zastosowaniu funkcji `log`. Wyniki są wyświetlone na rysunkach 6.12 i 6.13, co pozwala zilustrować różnicę. W funkcji `log` domyślnie używana jest podstawa 10. Aby zmniejszyć zbiór wyników i ułatwić generowanie wykresów, głębokość jest dodatkowo zaokrąglana za pomocą funkcji `round` do jednego miejsca po przecinku. Dane z tabeli są filtrowane, aby pominąć wyniki mniejsze niż 0,05, ponieważ po zaokrągleniu byłyby równe 0 lub mniej:

```
SELECT round(depth,1) as depth
,log(round(depth,1)) as log_depth
,count(*) as earthquakes
FROM earthquakes
WHERE depth >= 0.05
GROUP BY 1,2
;

depth  log_depth            earthquakes
-----  -------------------  -----------
0.1    -1.0000000000000000  6994
0.2    -0.6989700043360188  6876
0.3    -0.5228787452803376  7269
...    ...                  ...
```



Rysunek 6.12. Rozkład liczby trzęsień ziemi według głębokości bez dostosowywania jej poziomu



Rysunek 6.13. Rozkład liczby trzęsień ziemi według głębokości na skali logarytmicznej

Na rysunku 6.12 widać, że występuje duża liczba trzęsień ziemi na głębokości od 0,05 do mniej więcej 20. Dalej rozkład jest trudny do oceny, ponieważ oś *x* rozciąga się aż do 700, aby uwzględnić cały zakres danych. Jednak po przekształceniu głębokości na skalę logarytmiczną (rysunek 6.13) rozkład mniejszych wartości jest dużo łatwiejszy do analizy. Widoczny jest tu skok w punkcie 1,0, co odpowiada głębokości 10 km.

W SQL-u można też przeprowadzać inne rodzaje skalowania, choć nie zawsze nadają się one do eliminowania wartości odstających. Oto niektóre metody skalowania:

- ✓wyciąganie pierwiastka kwadratowego za pomocą funkcji `sqrt`;
- ✓wyciąganie pierwiastka trzeciego stopnia za pomocą funkcji `cube`;
- ✓przekształcanie na odwrotność (1 / `nazwa_pola`).

Aby zmienić jednostki, na przykład cale na stopy lub funty na kilogramy, należy pomnożyć lub podzielić wartość przez odpowiedni czynnik za pomocą operatorów * lub /.

Skalowanie można przeprowadzić w SQL-u, a także w oprogramowaniu lub języku używanym do generowania wykresów. Przekształcanie danych na skalę logarytmiczną jest przydatne zwłaszcza w sytuacji, gdy występuje duży rozrzut wartości dodatnich, a wzorce, których wykrycie jest istotne, występują w zakresie niskich wartości.

Podobnie jak we wszystkich analizach decyzja o tym, jak traktować anomalie, zależy od przeznaczenia zbioru danych, a także od wiedzy z dziedziny i znajomości kontekstu. Usuwanie wartości odstających to najprostsza technika, ale jeśli chcesz zachować wszystkie rekordy, możesz też posłużyć się winsoryzacją i zmianą skali.

Podsumowanie

Wykrywanie anomalii jest często potrzebne w trakcie analiz. Celem może być wykrywanie wartości odstających lub modyfikowanie ich, aby przygotować zbiór danych do dalszych analiz. W obu scenariuszach w skutecznym znajdowaniu anomalii pomagają podstawowe metody: sortowanie, obliczanie percentyli i wyświetlanie na wykresie danych wyjściowych z zapytań SQL-owych. Anomalie przyjmują wiele form. Najczęściej są to: wartości odstające, nietypowe skoki aktywności i nietypowe braki. Wiedza z danej dziedziny prawie zawsze pomaga w procesie szukania i zbierania informacji na temat przyczyn anomalii. Sposoby radzenia sobie z anomaliami to: badanie ich, usuwanie, zastępowanie innymi wartościami i skalowanie. Wybór metody zależy w dużym stopniu od celu analiz. Wszystkie te techniki można stosować za pomocą SQL-a. W następnym rozdziale przechodzę do eksperymentów, które mają prowadzić do stwierdzenia, czy cała grupa badana różni się od norm typowych dla grupy kontrolnej.

SQL:
tak wyciągniesz z danych rzetelne wnioski!

Język SQL został stworzony jako narzędzie do przetwarzania danych. Mimo że zwykle jest używany do pracy z bazami danych, jego możliwości są o wiele większe. Poprawny kod SQL ułatwia przetwarzanie potężnych zbiorów danych z dużą szybkością. Szczególnie obiecującą perspektywą jest zastosowanie języka SQL na wielkich zbiorach danych przechowywanych w chmurze. Dzięki nieco bardziej złożonym konstrukcjom SQL analityk danych może z dużą efektywnością wydobywać z nich wiedzę.

Ta praktyczna książka jest przeznaczona dla analityków danych i danologów, którzy chcą używać SQL-a do eksploracji dużych zbiorów danych. Pokazuje zarówno popularne, jak i nieco mniej znane techniki budowania zapytań SQL, dzięki czemu możliwe staje się rozwiązywanie nawet bardzo zawiłych problemów i optymalne wykorzystanie właściwości tego języka w pracy na danych. W nowy, innowacyjny sposób przedstawiono tu takie pojęcia jak złączenia, funkcje okna, podzapytania i wyrażenia regularne. Zademonstrowano, jak łączyć różne techniki, aby szybciej osiągać cele za pomocą łatwego do zrozumienia, czytelnego kodu. Opisywany materiał został zilustrowany licznymi przykładami zapytań SQL, dzięki czemu można płynnie przejść do rozwiązywania konkretnych problemów z zakresu przetwarzania, analizy i eksploracji danych.

Najciekawsze zagadnienia:

- przygotowywanie danych do analizy
- analizy szeregów czasowych z wykorzystaniem SQL
- analizy kohortowe do badania zachodzących zmian
- analiza tekstu za pomocą zaawansowanych funkcji i operatorów SQL
- wykrywanie odstających wartości
- analizy eksperymentów (testy A/B)

Cathy Tanimura jest analityczką danych z ponad dwudziestoletnim doświadczeniem. Odnosiła również sukcesy, budując zespoły do spraw analizy danych i tworząc potrzebną infrastrukturę. Zajmowała się także zarządzaniem zespołami w kilku czołowych firmach technologicznych. Od wielu lat używa języka SQL do pracy z większością komercyjnych i otwartych baz danych.

ISBN 978-83-283-8895-6

Cena: 69,00 zł