

Spis treści

Przedmowa	9
1. Analizy z wykorzystaniem SQL-a	13
Czym jest analiza danych?	13
Dlaczego SQL?	15
Czym jest SQL?	15
Korzyści, jakie daje SQL	18
SQL a R lub Python	19
SQL jako element procesu analizy danych	20
Rodzaje baz danych i sposoby pracy z nimi	22
Wierszowe bazy danych	23
Kolumnowe bazy danych	25
Inne rodzaje infrastruktury danych	26
Podsumowanie	27
2. Przygotowywanie danych do analiz	28
Typy danych	29
Typy danych w bazach	29
Dane ustrukturyzowane i nieustrukturyzowane	30
Dane ilościowe i jakościowe	31
Dane z pierwszej, drugiej i trzeciej ręki	32
Dane rzadkie	33
Struktura zapytań w SQL-u	33
Profilowanie — rozkład danych	36
Histogramy i częstość wystąpień	36
Binning	39
Technika n przedziałów	41
Profilowanie — jakość danych	43
Wykrywanie duplikatów	43
Deduplikacja za pomocą klauzul GROUP BY i DISTINCT	45

Przygotowania — oczyszczanie danych	46
Oczyszczanie danych za pomocą przekształceń w instrukcji CASE	46
Konwersja i rzutowanie typów	49
Radzenie sobie z wartościami null: funkcje coalesce, nullif i nvl	51
Brakujące dane	54
Przygotowania — kształtowanie danych	58
Docelowe zastosowanie — analiza biznesowa, wizualizacja, obliczanie statystyk, uczenie maszynowe	58
Tworzenie tabel przestawnych za pomocą instrukcji CASE	59
Przywracanie struktury po przestawieniu z użyciem instrukcji UNION	61
Funkcje pivot i unpivot	63
Podsumowanie	64
3. Analiza szeregów czasowych	66
Operacje na datach, czasie oraz datach i czasie	67
Zmiana strefy czasowej	67
Konwersja formatu dat i znaczników czasu	69
Obliczenia matematyczne na datach	72
Obliczenia na czasie	75
Złączanie danych z różnych źródeł	76
Zbiór danych o sprzedaży detalicznej	77
Analiza trendów w danych	77
Proste trendy	78
Porównywanie komponentów	80
Obliczanie procentów z całości	88
Stosowanie indeksacji do badania zmian procentowych w czasie	91
Okna przesuwne	95
Obliczenia na podstawie okien przesuwnych	97
Okna przestawne w rzadkich zbiorach danych	101
Obliczanie wartości skumulowanych	104
Analiza danych z efektem sezonowości	106
Porównywanie okres do okresu — rdr i mdm	107
Porównania okres do okresu — te same miesiące z kolejnych lat	109
Porównywanie z wieloma wcześniejszymi okresami	114
Podsumowanie	116
4. Analiza kohortowa	117
Kohorty — przydatny model analiz	117
Zbiór danych o członkach Kongresu	120
Utrzymanie	122
Kod w SQL-u do tworzenia prostej krzywej utrzymania	123

Modyfikowanie szeregów czasowych, aby zwiększyć dokładność wyników analizy utrzymania	126
Kohorty tworzone na podstawie szeregów czasowych	131
Definiowanie kohort na podstawie odrębnej tabeli	136
Jak radzić sobie z kohortami rzadkimi?	140
Definiowanie kohort na podstawie dat innych niż początkowa	144
Powiązane analizy kohortowe	146
Przeżywalność	146
Powroty (ponowne zakupy)	150
Obliczanie skumulowanych wartości	155
Analiza przekrojowa w kontekście analizy kohortowej	158
Podsumowanie	165
5. Analiza tekstu	166
Po co analizować tekst za pomocą SQL-a?	166
Czym jest analiza tekstu?	166
Dlaczego SQL jest dobrym narzędziem do analizy tekstu?	167
Kiedy SQL nie jest dobrym wyborem?	168
Zbiór danych o obserwacjach UFO	169
Cechy tekstu	170
Parsowanie tekstu	172
Przekształcanie tekstu	176
Znajdowanie elementów w większych blokach tekstu	183
Dopasowywanie symboli wieloznacznych: LIKE i ILIKE	184
Dokładne dopasowywanie za pomocą operatorów IN i NOT IN	188
Wyrażenia regularne	191
Tworzenie tekstu i zmienianie jego kształtu	204
Konkatencja	205
Zmiana kształtu tekstu	208
Podsumowanie	211
6. Wykrywanie anomalii	212
Możliwości i ograniczenia SQL-a w zakresie wykrywania anomalii	213
Zbiór danych	214
Wykrywanie wartości odstających	215
Wyszukiwanie anomalii za pomocą sortowania	215
Wyszukiwanie anomalii na podstawie percentyli i odchylenia standardowego	218
Tworzenie wykresów w celu znajdowania anomalii	224
Rodzaje anomalii	232
Anomalne wartości	232
Anomalne liczby wystąpień	235
Anomalie w postaci braku danych	239

Radzenie sobie z anomaliami	241
Badanie anomalii	241
Usuwanie danych	242
Zastępowanie innymi wartościami	243
Skalowanie	244
Podsumowanie	247
7. Analiza eksperymentów	248
Wady i zalety analizy eksperymentów za pomocą SQL-a	249
Zbiór danych	250
Rodzaje eksperymentów	252
Eksperymenty z wynikami binarnymi — test chi-kwadrat	252
Eksperymenty z wynikami ciągłymi — test t	254
Problemy z eksperymentami i sposoby radzenia sobie z błędami	256
Przydział jednostek do wariantów	256
Wartości odstające	257
Okna czasowe	258
Eksperymenty związane z wielokrotną ekspozycją	259
Co robić, gdy kontrolowane eksperymenty są niemożliwe? Inne analizy	261
Analiza „przed i po”	261
Analiza eksperymentów naturalnych	263
Analiza populacji w okolicy wartości progowej	264
Podsumowanie	265
8. Tworzenie złożonych zbiorów danych na potrzeby analiz	266
Kiedy używać SQL-a do tworzenia złożonych zbiorów danych?	266
Zalety stosowania SQL-a	267
Kiedy używać procesu ETL?	267
Kiedy umieszczać logikę w innych narzędziach?	268
Porządkowanie kodu	270
Komentarze	271
Wielkość liter, wcięcia, nawiasy i inne sztuczki z obszaru formatowania	272
Przechowywanie kodu	274
Porządkowanie obliczeń	274
Porządek przetwarzania klauzul w SQL-u	275
Podzapytania	278
Tabele tymczasowe	280
Wyrażenia CTE	281
Instrukcja grouping sets	282

Zarządzanie wielkością zbioru danych i prywatnością	285
Próbkowanie na podstawie wartości procentowych i dzielenia modulo	286
Zmniejszanie liczby wymiarów	287
Dane osobowe i prywatność danych	291
Podsumowanie	292
9. Podsumowanie	293
Analizy lejka	293
Rezygnacje, wygaśnięcia i inne definicje utraty klientów	294
Analiza koszykowa	298
Materiały	300
Książki i blogi	300
Zbiory danych	302
Uwagi końcowe	302