# 10

# FUNKCJE STATYSTYCZNE W SQL

Baza danych SQL zwykle nie jest pierwszym narzędziem wybieranym przez analityka danych podczas przeprowadzania analizy statystycznej, która wymaga czegoś więcej niż tylko obliczania sum i średnich. Typowo wybieranym oprogramowaniem byłyby pakiety statystyczne z wieloma funkcjonalnościami, takie jak SPSS lub SAS, języki programowania R lub Python, a nawet Excel. Jednak standardowy ANSI SQL, w tym implementacja PostgreSQL, oferuje garść potężnych funkcji statystycznych, które ujawnią wiele informacji o Twoich danych bez konieczności eksportowania zestawu danych do innego programu.

W tym rozdziale omówimy te funkcje statystyczne SQL wraz ze wskazówkami, kiedy ich używać. Statystyka jest obszernym tematem wartym opisania w odrębnej książce, więc przejrzymy go tutaj tylko pobieżnie. Niemniej jednak nauczysz się, jak stosować zaawansowane pojęcia statystyczne w celu poszukiwania informacji w nowym zestawie danych z amerykańskiego biura spisu ludności. Nauczysz się również używać SQL do tworzenia porównań z zastosowaniem rankingów na danych pochodzących z FBI na temat przestępstw.

## Tworzenie tabeli z danymi statystycznymi ze spisu ludności

Powróćmy do jednego z moich ulubionych źródeł danych, czyli amerykańskiego biura spisowego. W rozdziałach 4. i 5. zostały użyte dane ze spisu ludności z 2010 roku, aby je zaimportować i wykonać podstawowe operacje matematyczne i statystyczne. Tym razem użyjesz danych z hrabstw opracowanych na podstawie 5-letnich szacunków Amerykańskiego Badania Społecznego (ACS) na lata 2011–2015, oddzielnego badania przeprowadzonego przez Biuro ds. Spisu Ludności (Census Bureau).

Użyj kodu z listingu 10.1, aby utworzyć tabelę `acs_2011_2015_stats`, i zaimportuj plik CSV *acs_2011_2015_stats.csv*. Kod i dane są dostępne wraz ze wszystkimi innymi zasobami książki pod adresem *https://www.nostarch.com/practicalSQL/*. Pamiętaj, aby zmienić *C:\YourDirectory* na lokalizację zawierającą Twój plik CSV.

```
CREATE TABLE acs_2011_2015_stats (
  ❶ geoid varchar(14) CONSTRAINT geoid_key PRIMARY KEY,
    county varchar(50) NOT NULL,
    st varchar(20) NOT NULL,
  ❷ pct_travel_60_min numeric(5,3) NOT NULL,
    pct_bachelors_higher numeric(5,3) NOT NULL,
    pct_masters_higher numeric(5,3) NOT NULL,
    median_hh_income integer,
  ❸ CHECK (pct_masters_higher <= pct_bachelors_higher)
);
COPY acs_2011_2015_stats
FROM 'C:\YourDirectory\acs_2011_2015_stats.csv'
WITH (FORMAT CSV, HEADER, DELIMITER ',');
❹ SELECT * FROM acs_2011_2015_stats;
```

*Listing 10.1. Tworzenie tabeli dla danych statystycznych ACS oraz import danych*