Airlines Project

Łukasz Chrostowski

2023 Kwiecień

Opis projektu

Celem projektu jest przekrojowa analiza danych lotniczych związanych z opóźnienami lotów pochodzących USA i mierzonych w lipcu 2017 roku. Zadania zostały wykonane w ramch kursu `SQL w Analizie danych` na kierunku `Analiza i przetwarzanie danych` prowadzonego na Wydziale Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza. Projekt został podzielony na dwie części w zależności od typu zadań i analiz do przeprowadzenia. Niniejszy dokument przedstawia pierwszą z nich.

Zadania

Na wstępie warto załadować wymaganie pakiety do obsługi składni `SQL`. Następnie tworzymy obiekt do korzystania z bazy danych oraz pisania zapytań.

```
library(DBI)
library(RPostgres)
con <- dbConnect(RPostgres::Postgres(),
                 dbname = "dwbd_flights",
                 host = "psql.wmi.amu.edu.pl",
                 user = rstudioapi::askForPassword(prompt = "Database user"),
                 password = rstudioapi::askForPassword(prompt = "Database password")
)
```

Kilka statystyk dotyczących opóźnień lotów:

Jakie było średnie opóźnienie przylotu?

```
SELECT avg(arr_delay_new) AS avg_delay
FROM "Flight_delays";
```

Table 1: 1 records

avg_delay
15.91152

Jakie było maksymalne opóźnienie przylotu?

```
SELECT max(arr_delay_new) AS max_delay
FROM "Flight_delays";
```

Table 2: 1 records

max_delay
1895

Który lot miał największe opóźnienie przylotu?

```
SELECT carrier,
       origin_city_name,
       dest_city_name,
       fl_date,
       arr_delay_new
FROM "Flight_delays"
WHERE arr_delay_new IS NOT NULL
ORDER BY arr_delay_new DESC
LIMIT 1;
```

Table 3: 1 records

carrier	origin_city_name	dest_city_name	fl_date	arr_delay_new
AA	Kona, HI	Los Angeles, CA	2017-07-26	1895

Które dni tygodnia są najgorsze do podróżowania?

```
SELECT CASE day_of_week
                WHEN 1 THEN 'Monday'
                WHEN 2 THEN 'Tuesday'
                WHEN 3 THEN 'Wednesday'
                WHEN 4 THEN 'Thursday'
                WHEN 5 THEN 'Friday'
                WHEN 6 THEN 'Satrurday'
                WHEN 7 THEN 'Sunday'
        END AS weekday_name,
       avg(arr_delay_new) AS avg_delay
FROM "Flight_delays"
GROUP BY day_of_week
ORDER BY avg_delay DESC;
```

Table 4: 7 records

weekday_name	avg_delay
Friday	20.80747
Monday	18.04801
Wednesday	16.10514
Thursday	15.64696
Satrurday	15.21876
Tuesday	12.88056
Sunday	12.77606

Które linie lotnicze latające z San Francisco (SFO) mają najmniejsze opóźnienia przylotu?

```
SELECT DISTINCT T.airline_name,
       T.avg_delay
FROM
(
SELECT A.airline_name,
       avg(arr_delay_new) AS avg_delay
FROM "Flight_delays" F
    INNER  JOIN "Airlines" A
        ON A.airline_id = F.airline_id
WHERE A.airline_id IS NOT null
GROUP BY A.airline_name
) AS T
INNER JOIN "Airlines" A1
ON A1.airline_name = T.airline_name
INNER JOIN "Flight_delays" F1
ON F1.airline_id = A1.airline_id
WHERE F1.origin = 'SFO'
ORDER BY avg_delay DESC;
```

Table 5: Displaying records 1 - 10

airline_name	avg_delay
JetBlue Airways: B6	28.841148
Frontier Airlines Inc.: F9	18.980300
American Airlines Inc.: AA	18.375314
United Air Lines Inc.: UA	16.950403
SkyWest Airlines Inc.: OO	16.808273
Virgin America: VX	13.964467
Southwest Airlines Co.: WN	13.823983
Delta Air Lines Inc.: DL	12.258788
Alaska Airlines Inc.: AS	7.453927
Hawaiian Airlines Inc.: HA	4.202719

Jaka część linii lotniczych ma regularne opóźnienia, tj. jej lot ma średnio co najmniej 10 min. opóźnienia?

```
WITH
  T1 AS (
    SELECT COUNT(average_late) AS count10
    FROM (
      SELECT AVG(arr_delay_new) AS average_late
      FROM "Flight_delays"
      GROUP BY airline_id
      ) AS T
    WHERE T.average_late > 10
  ),
  T2 AS (
    SELECT COUNT(average_late) AS tot_count
    FROM (
```

```
      SELECT AVG(arr_delay_new) AS average_late
      FROM "Flight_delays"
      GROUP BY airline_id
      ) AS T
  )
SELECT
  CAST(T1.count10 AS FLOAT) / T2.tot_count AS late_proporstion
FROM
  T1,
  T2;
```

Table 6: 1 records

late_proporstion
0.8333333

Jak opóźnienia wylotów wpływają na opóźnienia przylotów?

```
SELECT (dep_arr_mean - arr_mean * dep_mean) / (arr_std * dep_std) AS "Pearson r"
FROM
(SELECT avg(arr_delay_new) AS arr_mean,
       avg(dep_delay_new) AS dep_mean,
       avg(arr_delay_new * dep_delay_new) AS dep_arr_mean,
       stddev(arr_delay_new) AS arr_std,
       stddev(dep_delay_new) AS dep_std
FROM "Flight_delays") AS T;
```

Table 7: 1 records

Pearson r
0.9737081

Która linia lotnicza miała największy wzrost (różnica) średniego opóźnienia przylotów w ostatnim tygodniu miesiąca, tj. między 1-23 a 24-31 lipca?

```
WITH
T1 AS
(
SELECT A.airline_name,
       avg(F.arr_delay_new) AS mean
FROM "Flight_delays" F
INNER JOIN "Airlines" A
ON A.airline_id = F.airline_id
WHERE day_of_month < 24
GROUP BY A.airline_name
), T2
AS
(
SELECT A.airline_name,
       avg(F.arr_delay_new) AS mean
```

```
FROM "Flight_delays" F
INNER JOIN "Airlines" A
ON A.airline_id = F.airline_id
WHERE day_of_month >= 24
GROUP BY A.airline_name
)
SELECT  T1.airline_name,
        T2.mean - T1.mean AS delay_increase
FROM T1
INNER JOIN T2
ON T1.airline_name = T2.airline_name
where T2.mean - T1.mean >= ALL (SELECT T2.mean - T1.mean
                                FROM T1
                                INNER JOIN T2
                                ON T1.airline_name = T2.airline_name);
```

Table 8: 1 records

airline_name	delay_increase
Southwest Airlines Co.: WN	0.584763

Które linie lotnicze latają zarówno na trasie SFO → PDX (Portland), jak i SFO → EUG (Eugene)?

```
WITH
T1 AS
(
SELECT DISTINCT(A.airline_name)
FROM "Flight_delays" F
INNER JOIN "Airlines" A
ON F.airline_id = A.airline_id
WHERE origin = 'SFO' AND dest = 'PDX'
), T2
AS
(
SELECT DISTINCT(A.airline_name)
FROM "Flight_delays" F
INNER JOIN "Airlines" A
ON F.airline_id = A.airline_id
WHERE origin = 'SFO' AND dest = 'EUG'
)
SELECT T1.airline_name
FROM T1
INNER JOIN T2
ON T1.airline_name = T2.airline_name;
```

Table 9: 2 records

airline_name
SkyWest Airlines Inc.: OO
United Air Lines Inc.: UA

Jak najszybciej dostać się z Chicago do Stanfordu, zakładając wylot po 14:00 czasu lokalnego?

```
SELECT origin,
       dest,
       avg(arr_delay_new) AS mean
FROM "Flight_delays"
WHERE crs_dep_time > 1400 AND (origin = 'MDW' OR origin = 'ORD') AND (dest = 'SFO' OR dest = 'SJC' OR d
GROUP BY origin, dest
ORDER BY avg(arr_delay_new) desc;
```

Table 10: 5 records

origin	dest	mean
ORD	SFO	22.19253
MDW	SFO	19.85714
MDW	SJC	17.20000
ORD	SJC	14.81111
MDW	OAK	12.12903