Tíðni nafna á Íslandi

Jakob Stefán Ívarsson

2024-09-23

1. Búið til töflu sem inni heldur gögn um eiginnöfn og millinöfn ásamt tíðni þeirra.

Fyrst þurfum við að tengjast við SQLite gagnagrunninn þar sem við munum geyma gögn um nöfn. Ef taflan „names“ er þegar til, munum við eyða henni. Eftir það gerum við nýja töflu sem inniheldur nafn, ár, tíðni og tegund nafna. Eftirfarandi R-kóða bútar eru nauðsynleg aðferð til að geta notað SQLite innan RStudio forritsins. Hér tengjumst við SQL og búum til Gagnasafn sem heitir “names_freq.db”. Þaðan af búum við til töflu sem heitir names, inn í SQL gagnagrunninum. Hún inniheldur 4 dálka: “name”, “year”, “frequency” og “type”. Við veljum síðan dálkana name, year og type sem primary key. Samsetning af þessum þremur dálkum sem primary key tryggir að hver færsla í töflunni sé einstök.
Dæmi:
Ef nafnið “Jón” er skráð sem eiginnafn árið 2000 og líka sem millinafn árið 2000, þá eru þetta tvær ólíkar færslur. Fyrir “Jón” árið 2010 myndi önnur færsla bætast við fyrir hvert ár og tegund. Samsetti lykillinn (name, year, type) tryggir að allar þessar færslur eru meðhöndlaðar sem einstakar færslur sem viðheldur gagnaheilleika (e. data integrity). Það er í þessu tilfelli er það betra heldur en að hafa bara einn primary key eins og nafn sem er ekki alveg nógu nákvæmt.

```
con <- dbConnect(RSQLite::SQLite(), "names_freq.db")

dbExecute(con, "DROP TABLE IF EXISTS names;")
```

```
## [1] 0
```

```
nafnatafla <- "
CREATE TABLE names (
    name TEXT NOT NULL,
    year INTEGER NOT NULL,
    frequency INTEGER NOT NULL,
    type TEXT NOT NULL,
    PRIMARY KEY (name, year, type)
);"

dbExecute(con, nafnatafla)
```

```
## [1] 0
```

Nú lesum við inn gögnin úr CSV skránum: first_names_freq.csv og middle_names_freq.csv (inniheldur gögn fyrir eiginnöfn og millinöfn). Við bætum við nýjum dálki í nýja gagnasafnið sem gefur til kynna tegund nafnsins, hvort það sé „eiginnafn“ eða „millinafn“. Eftir það sameinum við gögnin úr báðum skrám í eina

stærri töflu. Eftirfarandi aðferð var notuð í R til að geta framkvæmt verkefnið með SQL í RStudio. Þetta setur csv skrárnar saman undir nafninu sameina og inn í "names" töfluna í names_freq.db SQL gagnasafninu

```
first_names <- read.csv("data/first_names_freq.csv")
first_names$type <- "eiginnafn"


middle_names <- read.csv("data/middle_names_freq.csv")
middle_names$type <- "millinafn"

sameina <- bind_rows(
    first_names %>% rename(frequency = count),
    middle_names %>% rename(frequency = count)
)
dbWriteTable(con, "names", sameina, append = TRUE, row.names = FALSE)
```

2. Greining: Greinið tíðni nafna allra hópmeðlima teymsins út frá þessum gögnum. Notið eina SQL fyrirspurn til að svara hverju af eftirfarandi spurningum:

a) Hvaða hópmeðlimur á algengasta eiginnafnið?

Fyrir þessa spurningu notum við eina SQL fyrirspurn til að finna hópmeðliminn með mesta fjölda tíðni eiginnafns síns. Kóðinn hér að neðan skilar nafni hvers og eins, hámarks tíðni þann einstaklings og tegund nafnsins. Þessi SQL fyrirspurn finnur því hámarks tíðni (max_tidni) fyrir nöfnin 'Brynjar', 'Jakob' og 'Halldór', en einungis þar sem nafnið er skráð sem type, 'eiginnafn'. Að lokum raðar hún niðurstöðunum í töflu, raðað í stafrófsröð eftir nafni.

```
SELECT name,
       MAX(frequency) AS max_tidni,
       type
FROM names
WHERE name IN ('Brynjar', 'Jakob', 'Halldór') AND type = 'eiginnafn'
GROUP BY name, type
HAVING MAX(frequency) = (
    SELECT MAX(frequency)
    FROM names AS sub
    WHERE sub.name = names.name AND sub.type = 'eiginnafn'
)
ORDER BY name;
```

Table 1: 3 records

name	max_tidni	type
Brynjar	29	eiginnafn
Halldór	39	eiginnafn
Jakob	16	eiginnafn

Til að prenta EINUNGIS þann einstakling sem hefur hæstu nafnatíðnina getum við notað eftirfarandi SQL fyrirspurn:

```sql
SELECT name,
       frequency AS max_tidni,
       type
FROM names
WHERE name IN ('Brynjar', 'Jakob', 'Halldór')
  AND type = 'eiginnafn'
ORDER BY frequency DESC
LIMIT 1;
```

Table 2: 1 records

name	max_tidni	type
Halldór	39	eiginnafn

Halldór er klárlega algengasta nafnið með 39 tilfelli á einum tímapunkti.

b) Hvenær voru öll nöfnin vinsælust?

Hér notum við SQL fyrirspurn til að finna árið þegar nöfnin voru með hæstu tíðnina. Kóðinn skilar nafninu, árinu, hámarks tíðni og tegund nafnsins.

```sql
SELECT name,
       year,
       frequency AS max_tidni,
       type
FROM names
WHERE name IN ('Brynjar', 'Jakob', 'Halldór')
  AND (name, frequency) IN (
      SELECT name, MAX(frequency)
      FROM names
      WHERE name IN ('Brynjar', 'Jakob', 'Halldór')
      GROUP BY name
  )
ORDER BY name, type;
```

Table 3: 3 records

name	year	max_tidni	type
Brynjar	1998	29	eiginnafn
Halldór	1957	39	eiginnafn
Jakob	1998	16	eiginnafn

Hér sjáum við að vinsælasta árið fyrir Brynjar var 1998. Halldór naut mest vinsælda 1957 og Jakob 1998. Þessi ár voru þau öll vinsælust sem eiginnafn.

c) Hvenær komu nöfnin fyrst fram?

Fyrir þessa spurningu sköpum við SQL fyrirspurn sem skilar elsta tilfelli þ.e. ári sem nafn kom fyrst fram í sameinaða gagnagrunninum. Að auki höfðum við hvort nafnið hafi verið millinafn eða eiginnafn með í niðurstöðurnar.

```sql
SELECT name, year AS fyrst_fram, type
FROM names
WHERE (name, year) IN (
    SELECT name, MIN(year) AS min_year
    FROM names
    WHERE name IN ('Brynjar', 'Jakob', 'Halldór')
    GROUP BY name
)
ORDER BY name;
```

Table 4: 4 records

name	fyrst_fram	type
Brynjar	1927	millinafn
Halldór	1908	eiginnafn
Jakob	1908	eiginnafn
Jakob	1908	millinafn

Sjáum að elsta ummerki um Brynjar er sem millinafn árið 1927. Halldór var notað fyrst sem eiginnafn árið 1908. Sama má segja um Jakob en þar var elsta ummerki um Jakob bæði eiginnafn og millinafn árið 1908.