Tecnológico de Monterrey

SQL Project: Knowing Mexican Population

Team members with TEC ids A01632840, A01638214, A01632202, A01351361, A00227490, A01637876, A01637592

May 18th 2022

Advanced Databases

Carlos Emiliano Brito Nieto

Carlos Estrada Ceballos

Eduardo Esteva Camacho

Erick Gerardo Calderón Reyes

Ernesto Adrián Álvarez Salazar

Jessica Nicole Copado Leal

Samuel Alejandro Díaz del Guante Ochoa

Tecnológico de Monterrey

Team members
Advanced Databases Miguel Angel Palomera Pérez.

Census of Population and Housing 2020

The Census of Population and Housing 2020 (Census of 2020) was carried out from March 2nd to the 27th. A little more than 147 thousand interviewers participated in it, and they covered nearly two million square kilometers of the national territory, visiting each of the housing units to obtain information about them, count the population living in Mexico and inquire about their main demographic, socioeconomic and cultural characteristics.

Based on the information for the above census. Answer the following questions.

1. What is the median of kids per family grouped by population type (rural, urban)? Take into account that a population is rural if the number of habitants is less than 15000.
2. How many people speak a language other than spanish?
3. What is the median of the number of people per house grouped by state?
4. Calculate the number of people with bachelor degrees or more grouped by state. Present the results in a cross table (pivot table). The categories are the type of education.

Strategy

For importing the CSV, we decided to find the Columns that are the most imperative for getting information of the questions specifically, this reduced the size of the CSV files and allowed for faster search times.
If information was not specified from the CSV, we decided to define it ourselves in the tables as an added column, such as the type of population (rural, urban) according to the specified data given.

csvkit is a tool that can help to cut CSV columns so you only need those. After installing, use the following command to get the columns you need:

```
#: csvcut - c nameColumn[,nameColumn,nameColumn] nameOfFile.csv > newFile.csv
```

Database model

Model for Vivienda

```
CREATE TABLE vivienda (
   ID_VIV VARCHAR(12) PRIMARY KEY,
   ENT VARCHAR(2),
   NUMPERS INT,
   TAMLOC VARCHAR(1)
);

COPY vivienda(ENT, ID_VIV, NUMPERS, TAMLOC)
FROM '/workspace/SQL-Practice/CSV_poblacion/cpv2020_cb_viviendas_ej_Modified.csv'
DELIMITER ','
CSV HEADER;
```

id_viv	ent	numpers	tamloc
…	…	…	…

```
ALTER TABLE vivienda ADD COLUMN POP_TYPE VARCHAR;
```

id_viv	ent	numpers	tamloc	pop_type
…	…	…	…	…

Model for Personas

```
CREATE TABLE personas (
        ENT varchar(2),
        ID_VIV varchar(17),
        ID_PERSONA varchar(12) PRIMARY KEY,
        HLENGUA varchar(1),
        NIVACAD varchar(2),
        EDAD int
);
```

ENT	ID_VIV	ID_PERSONA	HLENGUA	NIVACAD	EDAD
…	…	…	…	…	…

```
COPY personas
FROM '/workspace/sql_practices/Censo2020_CPV_CB_Personas_ejemplo_csv.CSV'
WITH (FORMAT CSV, HEADER);
```

Queries

Question 1:

```
UPDATE vivienda
SET pop_type = 'rural'
WHERE tamloc = '2' OR tamloc = '1';

UPDATE vivienda
SET pop_type = 'urban'
WHERE tamloc = '3' OR tamloc = '4' OR tamloc = '5';

-- Identify all minors in every house (innermost query), then count the amount of minors per house, finally join and group by population type (rural or urba) and get median

SELECT PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY person.edad_count) AS median,
pop_type
FROM vivienda as viv
INNER JOIN (
   SELECT  COUNT(q1.edad) as edad_count, q1.id_viv
                  FROM (
                        SELECT id_persona, id_viv, edad
                        FROM personas
                        WHERE edad < 18 ORDER BY id_viv asc) as q1
   GROUP BY q1.id_viv) as person
ON viv.id_viv = person.id_viv
GROUP BY pop_type
```

Question 2:

```
select count(id_persona) as num_people_speak_other_than_spanish
from personas
where hlengua = '1';
```

Question 3:

```
SELECT PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY numpers) AS median,
ent
```

```
FROM vivienda
GROUP BY ent;
```

Question 4:

```
CREATE EXTENSION tablefunc;

SELECT *
from crosstab(
	$$
	SELECT ent, nivacad, COUNT(*)
	FROM personas
	WHERE nivacad = '11' OR
		nivacad = '12' OR
		nivacad = '13' OR
		nivacad = '14'
	GROUP BY ent, nivacad
	$$
)
AS (ent varchar,
	bachelor bigint,
	specialty bigint,
	master bigint,
	doctor bigint);
```

Results

1. **Result**: The median of kids per family in a rural setting is 3 kids per family, while on a urban setting, it's 2 kids per family

	median double precision	pop_type character varying
1	3	rural
2	2	urban

2. **Result:** 73,418 people speak a language other than spanish

	num_people_speak_other_than_spanish bigint
1	73418

3. **Result**: The median is 3 for both states 14 and 21, which are Jalisco and Puebla, respectively

	median double precision	ent character varying (2)
1	3	14
2	3	21

4. **Result:** The total number of people with bachelor degrees or higher by state (Jalisco and Puebla respectively) categorized by the types of education.

	ent character varying	bachelor bigint	specialty bigint	master bigint	doctor bigint
1	14	95022	2382	8343	1462
2	21	63542	1582	5617	1058

Links

- https://www.inegi.org.mx/programas/ccpv/2020/#Microdatos
- https://github.com/mpalomera/practical-sql-2