Estimating Housing Growth by Decade with SQL Queries

Lee Hachadoorian

Overview

Because census tracts change over time, comparisons of population change can be difficult. Housing data has a year built variable that can be used to estimate where housing units *were* located based on *current* Census data. More complex analyses (which will not be undertaken in this exercise) can combine housing data with population estimates to more precisely locate historical population in geographic units that are comparable across time.

Beging by loading all necessary libraries:

```
library(tidyverse)
library(tidycensus)
library(stringr)
library(sf)
library(tmap)
```

Data

We will be using data from two sources:

1. Tract-level data of the year a housing structure was built from ACS table B25036.
2. Time-series data of county-level counts of housing structures from NHGIS.

The exercise assumes the use of data for the Pennsylvania. You can choose a different state, or choose a group of nearby counties.

Use the `get_acs` function from the tidycensus package to download the data for table B25036. Complete the following code, referring to the help or previous exercises to fill out the `get_acs` function:

```
# Vector of variables from ACS table B25036-Tenure by Year Structure Built
year_built = paste("B25036", str_pad(1:23, 3, pad = "0"), sep = "_")
```

```
# Download year_built variables by tract for ACS 2015, including geometries
sfTracts = get_acs(...)
```

ACS Table B25036 has the following structure:

- B25036001 — Total:
- B25036002 — Owner occupied:
- B25036003 — Built 2010 or later
- B25036004 — Built 2000 to 2009
- B25036005 — Built 1990 to 1999
- B25036006 — Built 1980 to 1989
- B25036007 — Built 1970 to 1979
- B25036008 — Built 1960 to 1969
- B25036009 — Built 1950 to 1959
- B25036010 — Built 1940 to 1949
- B25036011 — Built 1939 or earlier
- B25036012 — Renter occupied:
- B25036013 — Built 2010 or later
- B25036014 — Built 2000 to 2009
- B25036015 — Built 1990 to 1999
- B25036016 — Built 1980 to 1989
- B25036017 — Built 1970 to 1979
- B25036018 — Built 1960 to 1969
- B25036019 — Built 1950 to 1959
- B25036020 — Built 1940 to 1949
- B25036021 — Built 1939 or earlier

The county data must be downloaded from NHGIS.org. As you used NHGIS data during the first lab exercise, no instructions are provided. Set Geographic Level to county and the Topic to "Total Housing Units". Then, make sure to select the Time Series Tables tab, which contains data for counties over a multi-decade period. Choose the version with "Nominal" integration (which means it treats counties as consistent units, even if there were border changes).

The data will be downloaded as a CSV (text) file. It will contain the following data columns:

- A41AA1970 — 1970: Housing units: Total
- A41AA1980 — 1980: Housing units: Total
- A41AA1990 — 1990: Housing units: Total
- A41AA2000 — 2000: Housing units: Total
- A41AA2010 — 2010: Housing units: Total

Use the `read_csv` function to read the data in. Rename the columns to make them easier to work with:

```
filename = "..."
dfCounties = read_csv(filename)
```

```
dfCounties = transmute(
  dfCounties, county_fips = paste(STATEFP, COUNTYFP, sep = ""),
  county_1970 = A41AA1970, county_1980 = A41AA1980, county_1990 = A41AA1990,
  county_2000 = A41AA2000, county_2010 = A41AA2010
)
```

Calculating Tract-Level Housing

Estimating Tract-Level Housing Based on Year Structure Built

ACS Table B25036 contains columns indicating the year that a residential structure was built. Therefore, we can estimate the number of housing units existing in a particular census year, by adding the columns for structures built in each of the preceding decades. Since the columns are split into owner occupied and renter occupied units, we have to add both of these together.

1970 is the earliest year for which the NHGIS time-series provides county-level totals. To calculate the number of structures existing in 1970, we need to add the categories `Built 1939 or earlier`, `Built 1940 to 1949`, `Built 1950 to 1959`, and `Built 1960 to 1969`, for both owner and occupied unites. Note the result of the following output:

```
transmute(
  sfTracts, GEOID,
  tract_est_1970 = B25036_011E + B25036_010E + B25036_009E + B25036_008E + B25036_021E + B25
)
```

As identifiers we will include the `GEOID` and a five digit county FIPS code constructed from the `STATEFP` and `COUNTYFP` fields. (We will use this later to join to the county housing data downloaded from NHGIS.)

Now edit the code to calculate columns representing the estimated housing in 1980 (`tract_est_1980`), 1990, 2000, and 2010. Note that you can immediately use newly named variables. Since I named a variable `tract_est_1970`, I can use this variable (instead of the 8 table columns that went into it) as input to create the `tract_est_1980` variable. In each case you will use the calculated column from the previous decade, and add two new fields, representing housing built up to the *prior* decade. For 1980, you will have to add `B25036007E` (representing owner-occupied housing built in the 1970s) and `B25036017E` (representing renter-occupied housing built in the 1970s).

```
# Calculate Decadal Housing Units
sfTracts = transmute(
  sfTracts, GEOID, NAME,
  county_fips = substr(GEOID, 1, 5),
```

```
  tract_est_1970 = B25036_011E + B25036_010E + B25036_009E + B25036_008E + B25036_021E + B25
  tract_est_1980 = tract_est_1970 + B25036_007E + B25036_017E,
  tract_est_1990 = ...,
  tract_est_2000 = ...,
  tract_est_2010 = ...
)
```

Open `sfTracts` in RStudio's data viewer to see the values that you calculated.

Adjusting Tract-Level Housing Estimate Based on County-Level Totals

Housing units from previous decades may have been demolished, in which case they won't appear in the current ACS data. On the other hand, units from previous decades may have been subdivided. Therefore, relying just on the year built data can lead to either an undercount or overcount of housing units from previous decades.

A simple way to adjust the count is to use the county-level housing counts from previous censuses, and adjust the tract-level count proportionally. The estimates calculated in the previous step are aggregated to the county level and compared to the official county-level housing unit count. If the aggregated estimates show 100,000 housing units in 1980 and the official county count is 110,000, then each tract estimate is multiplied by 1.1 (110,000/100,000).

Calculating County-Level Estimate Based on Year Structure Built

`sfTracts` now has tract-level estimates of housing units based on the year a structure was built. We can calculate an estimate of the county-level housing units by aggregating the tract-level estimates.

In order to do this, we are going to **group by** counties (using the `county_fips` column we constructed earlier) and use an aggregate operation, such as summation, averaging, counting, etc. In this case, we will find all tracts in the same county, and calculate the sum of each year's housing count. We store the result in a new data frame named `dfTracts`. Edit the following code to calculate the sum for each decade:

```
dfTracts = as.data.frame(sfTracts) %>%
  group_by(county_fips) %>%
  transmute(
    GEOID,
    county_est_1970 = sum(tract_est_1970),
    county_est_1980 = ...,
    county_est_1990 = ...,
    county_est_2000 = ...,
```

```
    county_est_2010 = ...
    ) %>%
  ungroup() %>%
  select(-county_fips)
```

View the resulting `dfTracts`.

Applying Proportional Adjustment to Tract-Level Estimates

For the final step, we need to adjust the tract-level estimate based on the county-level undercount or overcount. You should have already loaded the county-level data into the `dfCounties` data frame. For each year, the adjusted tract-level housing count is $tract_estimate * county_actual / county_estimate$.

Before we can calculate this, we need to join the data from three sources. `sfTracts` and `dfTracts` need to be joined based on a common field, then `dfCounties` needs to be joined on a different common field. See if you can figure out how to complete the two `inner_join` functions to accomplish this. Then complete the equations to calculate the adjusted housing for each decade:

```
# JOINING AND CALCULATING THE ADJUSTMENT VALUE
x = sfTracts %>%
  # Inner join the dfTracts table, which has the county estimates
  inner_join(...) %>%
  # Inner join the dfCounties table, which has the actual county housing counts
  inner_join(...) %>%
  # Calculate adjusted tract level housing counts
  mutate(
    tract_adj_1970 = tract_est_1970 * county_1970 / county_est_1970,
    tract_adj_1980 = ...,
    tract_adj_1990 = ...,
    tract_adj_2000 = ...,
    tract_adj_2010 = ...
    )
```

View the resulting object.

Mapping the Result

Build a map using tmap. You can pass in multiple variables, such as the 1970 and 2010 housing, to get "small multiples" or a series of related maps. Depending on the size of your map, you might want to filter `x` to show a smaller area. For example, if you downloaded all the data for Pennsylvania, you could use `filter(x, county_fips = "42101")` to show only tracts in Philadelphia:

```r
mapHousing = tm_shape(filter(x, county_fips = "42101")) +
  tm_fill(
    col = c("tract_adj_2010", "tract_adj_1970"),
    palette = "Blues", style = "jenks",
    title=expression("Urban Growth")
    )
mapHousing

save_tmap(tm = mapHousing, filename="FIGURE_NAME.jpeg")
```

ASSIGNMENT

Create a map showing the change between two time periods of your choice. For a choropleth map, choose a Graduated style. Set the column to an expression that converts the growth to a percentage. For example, to show growth between 1970 and 1990, you would need to create column using the following expression:

```
100 * (tract_adj_1990 - tract_adj_1970) / tract_adj_1970
```