Minnesota Population Center

Training and Development

IPUMS - NHGIS Extraction and Analysis

Exercise 2

OBJECTIVE: Gain an understanding of how the NHGIS datasets are structured and how they can be leveraged to explore your research interests. This exercise will use NHGIS datasets to explore changes in the number of college graduates living in Minnesota cities.

11/13/2017

Research Question

Which cities in Minnesota saw the greatest change in the number of collegeâ€¢educated residents since 1990?

Objectives

- Create and download an extract of NHGIS time series data
- Unzip data file and open in Microsoft Excel
- Analyze the data using Microsoft Excel
- Validate data analysis work using answer key

R Code to Review

This tutorial's sample code and answers use the so-called "tidyverse" style, but R has the blessing (and curse) that there are many different ways to do almost everything. If you prefer another programming style, please feel free to use it. But, for your reference, these are some quick explanations for commands that this tutorial will use:

- %>% - The pipe operator which helps make code with nested function calls easier to read. When reading code, it can be read as "and then". The pipe makes it so that code like `ingredients %>% stir() %>% cook()` is equivalent to `cook(stir(ingredients))` (read as "take *ingredients* and then *stir* and then *cook*").
- **summarize** - Summarize a datasets observations to one or more groups
- **group_by** - Set the groups for the summarize function to group by
- **filter** - Filter the dataset so that it only contains these values
- **mutate** - Add on a new variable to a dataset

Download Extract from IPUMS Website

Step 1: Log in to NHGIS

- Go to http://www.nhgis.org and click on â€˜Loginâ€™ in the top right.
- If you have already registered on any Minnesota Population Center websiteâ€¦
  - If you remember your password, log in now. Otherwise, click the â€œForgot your password?â€¢ link on the right and follow the instructions.
- If you have not already registeredâ€¦
  - Click on the â€œCreate an accountâ€¢ link on the right, fill in the required information, and submit your registration.

- You will then enter the NHGIS Data Finderâ€¦

Step 2: Investigate the Scope of Relevant Data

A common first step is to look into the range of data available on the topic of interest...

- Click the Topics filter button, then select â€˜Educational Attainmentâ€™, and submit the selection.

1) How many source tables are available for this topic? ___________

2) From what year is the oldest table that gives population counts by educational attainment? ___________

Step 3: Find Data for the Period of Interest

- With the topic already selected, click the Years filter button, then select â€˜1990â€™, and submit the selection.

The Select Data grid now lists all the tables related to the topic of â€˜Educational Attainmentâ€™ with data from 1990. One way to proceed would be to select one of the â€˜source tablesâ€™ listed here and then look for another more recent table to compare with it. However, the categories, terms, and universes used by census tables often change over time, which can make it difficult to pull together comparable data.

For many topics (including this one, conveniently!), NHGIS provides a simpler alternative: â€˜time series tablesâ€™, which link together comparable data from multiple years in one table.

- Click on the Time Series Table tab, located just right of the Source Tables tab at the top of the Select Data grid.
- Locate the following Time Series Table and answer the questions that follow:
  - Persons 25 Years and Over by Educational Attainment [7]

Learn About the Table in the Data Finder

3) Click the table name to see additional information. How many time series does this table contain?

____

4) Which 3 source tables are used to create this 1 time series table? ________________

____

5) What advantage is there in using this table rather than the â€˜Persons 18 Years and Over by Educational Attainment [7]â€™?

____

____

6) What type of â€˜geographic integrationâ€™ does this table use?
____

7) In the Select Data grid, click on â€˜Nominalâ€™ in the Geographic Integration column. With this type of integration, what should we keep in mind as we compare data across time? ____
____

Step 4: Create a Data Extract

Creating a data extract requires the user to select the table(s), specify a geographic level, and select the data layout structureâ€¦

- Click the plus sign to the left of the table name to add it to your Data Cart.
- Click the green Continue button under your Data Cart.
- On the Data Options screen, select the â€˜Placeâ€™ geographic level.
  - (In census terminology, cities, villages, and town centers are all â€˜placesâ€™.)
- Click the green Continue button under your Data Cart.
- On the Review and Submit screen:
  - Select the "Comma delimited (best for GIS)" option (it doesn't matter if you include the descriptive header rows or not)
  - Select "Time varies by row" (This is easiest to work with in R)
  - Add an extract description if you wish
  - Click Submit

Getting the data into R

Step 1: Download the Data Extract

From the Extracts History page, you will be able to download your data extract once it has finished processing, typically within a few minutes. You may leave this page and return once you have received the email alerting you to your finished extract.

If you refresh your browser window (click on the loop icon at top, or press F5), you will see the extract status change from â€˜queuedâ€™ to â€˜in progressâ€™ to â€˜completeâ€™, at which time you will be able to click the â€˜tablesâ€™ link to download the data.

- Return to the Extracts History page if not currently there.
- Rightâ€◎click on the â€˜tablesâ€™ link for the extract you created.
- Choose â€˜Save Target As...â€™ (or â€˜Save link as...â€™).

- Save the zip file into â€˜Documentsâ€™.
- The R package can read the extracts as zip files, or if you wish to open in other programs, you can unzip them, by: Rightâ€•clicking on the 'nhgis0001_csv.zip' file, and select Extract All... Then click the Extract button.

Step 2: Getting the data into R

You will need to change the filepaths noted below to the place where you have saved the extracts.

```
# Change these filepaths to the filepaths of your downloaded extract
nhgis_csv_file <- "nhgis0001_csv.zip"
nhgis_shp_file <- "nhgis0001_shape.zip"

library(ipumsr)
nhgis_ddi <- read_ipums_codebook(nhgis_csv_file) # Contains metadata, nice to
have as separate object
nhgis <- read_nhgis(nhgis_csv_file, verbose = FALSE)
```

This tutorial will also rely on the dplyr package, so if you want to run the same code, run the following command (but if you know other ways better, feel free to use them):

```
library(dplyr)
```

Part I: Analyze the Data

8) How many places are included in this table? ___________________________________

```
length(unique(nhgis$NHGISCODE))
```

9) Why do you think some places have missing values for some years? ____________

____

```
nhgis %>%
  group_by(NHGISCODE) %>%
  summarize(NAME = NAME[1], num_years = n())
```

10) How many place records are there for Minnesota? ____________________________
(Future questions will refer to the Minnesota subset)

```
mn <- nhgis %>%
  filter(STATE == "Minnesota")

length(unique(mn$NHGISCODE))
```

11) Aiming to compare counts of college graduates from 1990 and 2008â€•2012, it will be helpful first to think about only the columns of interest. Defining â€˜college graduatesâ€™ as anyone with a bachelorâ€™s degree or higher, which columns should we use? Note: The 2008â€•2012 data include both estimates and margins of error columns. For now, weâ€™re only interested in the estimate.

____

____

```
nhgis_ddi %>%
  ipums_var_info() %>%
  select(var_name, var_label) %>%
  filter(grepl("^B85", var_name) & !grepl("^Margin of error", var_label))

table(is.na(nhgis$B85AF), nhgis$YEAR)
table(is.na(nhgis$B85AG), nhgis$YEAR)
```

Create a new variables called â€œCollegeGradâ€•, and sum the appropriate counts to create totals for all places.

12) How many college graduates were living in White Bear Lake in 1990? __________

```
mn <- mn %>%
  mutate(CollegeGrad = B85AF + B85AG)

mn %>%
  select(NHGISCODE, PLACE, YEAR, CollegeGrad) %>%
  filter(grepl("^White Bear Lake", PLACE))
```

Summarize the table to calculate â€˜ChangeCollegeGradâ€™, and compute the total change in college grads between 1990 and 2008â€•2012 for all places.

13) Which city had the highest increase? How much was it? ______________________

```
mn_change <- mn %>%
  filter(YEAR %in% c("1990", "2008-2012")) %>%
  group_by(NHGISCODE) %>%
  filter(n() == 2) %>% # Only places available for both years
  # Convert to negative for 1990 so we can add them
  mutate(CollegeGrad = ifelse(YEAR == "1990", -CollegeGrad, CollegeGrad)) %>%
  summarize(PLACE = PLACE[1], ChangeCollegeGrad = sum(CollegeGrad))

mn_change %>%
  top_n(5, ChangeCollegeGrad) %>%
  arrange(desc(ChangeCollegeGrad))
```

We would expect that cities with great increases also had high overall population growth and vice versa. Continue working through the next set of questions if youâ€™d like to find out which cities had the greatest increases in the proportion of the population with bachelorâ€™s degrees.

Part II: College Grads by Place (Optional)

Create a new variable called Total, and sum the appropriate counts to get the total of all persons 25 years and over.

14) What was the total population 25+ of St. Paul in 2008â€•2012?

____

```
mn <- mn %>%
  mutate(
    TotalPop = B85AA + B85AB + B85AC + B85AD + B85AE + B85AF + B85AG
  )

mn %>%
  select(NHGISCODE, YEAR, PLACE, TotalPop) %>%
  filter(YEAR == "2008-2012" & grepl("St. Paul", PLACE))
```

Create a new variables called PctCollege. Multiply 100 times each CollegeGrad variable divided by each Total variable to calculate the percentage of the 25+ population with college degrees.

15) Which city had the highest percentage of college grads in 2008â€•2012?____________

Create a summary table with ChangePctCollege and calculate the differences between the PctCollege variables between 1990 and 2008â€•2012

```
mn <- mn %>%
  mutate(PctCollegeGrad = CollegeGrad / TotalPop * 100)

mn %>%
  filter(YEAR == "2008-2012") %>%
  select(NHGISCODE, PLACE, PctCollegeGrad, TotalPop, CollegeGrad) %>%
  top_n(5, PctCollegeGrad) %>%
  arrange(desc(PctCollegeGrad))
```

16) Which city had the highest increase in its proportion of college graduates?

____

```
mn_change <- mn %>%
  filter(YEAR %in% c("1990", "2008-2012")) %>%
  group_by(NHGISCODE) %>%
  filter(n() == 2) %>% # Only places available for both years
  # Convert to negative for 1990 so we can add them
  mutate(PctCollegeGrad = ifelse(YEAR == "1990", -PctCollegeGrad,
PctCollegeGrad)) %>%
  summarize(PLACE = PLACE[1], ChangeCollegeGrad = sum(PctCollegeGrad))

mn_change %>%
  top_n(5, ChangeCollegeGrad) %>%
  arrange(desc(ChangeCollegeGrad))
```

Answers

1) How many source tables are available when you filter only on Topic = â€˜Educational Attainmentâ€™?
953

2) From what year is the oldest table that gives population counts by educational attainment?
1934 â€“ (The 1880 table that appears for this topic has a universe of â€œschoolsâ€ and therefore does not provide â€œpopulation countsâ€ by educational attainment.)

3) How many time series does this table contain?
7

4) Which 3 source tables are used to create this 1 time series table?
NP57 from 1990 STF3, NP037C from 2000 SF 3a and B15002 from 2012 ACS 5â€Year

5) What advantage is there in using this table rather than the â€˜Persons 18 Years and Over by Educational Attainment [7]â€™?
A large portion of people aged 18â€24 are still actively working to complete a degree. The 25+ table helpfully captures the population after most have completed their formal education.

6) What type of â€˜geographic integrationâ€™ does this table use?
Nominal

7) With this type of integration, what should we keep in mind as we compare data across time?
This table wonâ€™t tell us how much of a cityâ€™s population changes were due to boundary changes, such as through annexation. Also, a city that changed its name or merged with another (e.g., Norwood Young America, MN, in 1997) will be missing values for some years.

8) How many places are included in this table?
30,544

```
length(unique(nhgis$NHGISCODE))
#> [1] 30544
```

9) Why do you think some places are missing values for certain years?
Possibilities: They didnâ€™t exist yet or ceased to exist at some point. They were unincorporated places that the Census did not identify in some years. The city changed its name or merged with another.

```
nhgis %>%
  group_by(NHGISCODE) %>%
  summarize(NAME = NAME[1], num_years = n())
#> # A tibble: 30,544 x 3
#>    NHGISCODE                 NAME num_years
#>        <chr>                <chr>     <int>
#>  1 G01000100 Abanda CDP, Alabama         1
#>  2 G01000124      Abbeville city         3
#>  3 G01000460     Adamsville city         3
#>  4 G01000484        Addison town         3
#>  5 G01000676          Akron town         3
#>  6 G01000820      Alabaster city         3
#>  7 G01000988    Albertville city         3
#>  8 G01001132 Alexander City city         3
#>  9 G01001180      Alexandria CDP         2
#> 10 G01001228     Aliceville city         3
#> # ... with 30,534 more rows
```

10) How many place records are there for Minnesota? *916*

```
mn <- nhgis %>%
  filter(STATE == "Minnesota")

length(unique(mn$NHGISCODE))
#> [1] 916
```

11) Defining â€˜college graduatesâ€™ as anyone with a bachelorâ€™s degree or higher, which columns should we highlight? *â€˜Bachelorâ€™s degreeâ€™ for both years and the â€˜Graduate or professional degreeâ€ for both years*

```
nhgis_ddi %>%
  ipums_var_info() %>%
  select(var_name, var_label) %>%
  filter(grepl("^B85", var_name) & !grepl("^Margin of error", var_label))
#> # A tibble: 7 x 2
#>   var_name
#>      <chr>
#> 1    B85AA
#> 2    B85AB
#> 3    B85AC
#> 4    B85AD
#> 5    B85AE
#> 6    B85AF
#> 7    B85AG
#> # ... with 1 more variables: var_label <chr>

table(is.na(nhgis$B85AF), nhgis$YEAR)
#>
#>          1990  2000 2008-2012
#>   FALSE 23435 25150     29509
```

```r
table(is.na(nhgis$B85AG), nhgis$YEAR)
#> 
#>          1990  2000 2008-2012
#>   FALSE 23435 25150     29509
```

12) How many college graduates were living in White Bear Lake in 1990?
    4,445

```r
mn <- mn %>%
  mutate(CollegeGrad = B85AF + B85AG)

mn %>%
  select(NHGISCODE, PLACE, YEAR, CollegeGrad) %>%
  filter(grepl("^White Bear Lake", PLACE))
#> # A tibble: 3 x 4
#>   NHGISCODE                PLACE      YEAR CollegeGrad
#>       <chr>                <chr>     <chr>       <int>
#> 1 G27069970 White Bear Lake city      1990        4445
#> 2 G27069970 White Bear Lake city      2000        4931
#> 3 G27069970 White Bear Lake city 2008-2012        5362
```

13) Which city had the highest increase? How much was it?
    Minneapolis: +40,568

```r
mn_change <- mn %>%
  filter(YEAR %in% c("1990", "2008-2012")) %>%
  group_by(NHGISCODE) %>%
  filter(n() == 2) %>% # Only places available for both years
  # Convert to negative for 1990 so we can add them
  mutate(CollegeGrad = ifelse(YEAR == "1990", -CollegeGrad, CollegeGrad)) %>%
  summarize(PLACE = PLACE[1], ChangeCollegeGrad = sum(CollegeGrad))

mn_change %>%
  top_n(5, ChangeCollegeGrad) %>%
  arrange(desc(ChangeCollegeGrad))
#> # A tibble: 5 x 3
#>   NHGISCODE            PLACE ChangeCollegeGrad
#>       <chr>            <chr>             <int>
#> 1 G27043000 Minneapolis city             40568
#> 2 G27058000    St. Paul city             20224
#> 3 G27071428    Woodbury city             17214
#> 4 G27054880   Rochester city             15535
#> 5 G27051730    Plymouth city             14519
```

14) What was the total population 25+ of St. Paul in 2008â€•2012?
    174,459

```r
mn <- mn %>%
  mutate(
    TotalPop = B85AA + B85AB + B85AC + B85AD + B85AE + B85AF + B85AG
  )
```

```
mn %>%
  select(NHGISCODE, YEAR, PLACE, TotalPop) %>%
  filter(YEAR == "2008-2012" & grepl("St. Paul", PLACE))
#> # A tibble: 5 x 4
#>   NHGISCODE      YEAR               PLACE TotalPop
#>       <chr>     <chr>               <chr>    <int>
#> 1 G27047221 2008-2012 North St. Paul city     7724
#> 2 G27058000 2008-2012       St. Paul city   174459
#> 3 G27058018 2008-2012  St. Paul Park city     3612
#> 4 G27061492 2008-2012 South St. Paul city    13928
#> 5 G27069700 2008-2012  West St. Paul city    13937
```

15) Which city had the highest percentage of college grads in 2008â€•2012?
Woodland: 79.8%

```
mn <- mn %>%
  mutate(PctCollegeGrad = CollegeGrad / TotalPop * 100)

mn %>%
  filter(YEAR == "2008-2012") %>%
  select(NHGISCODE, PLACE, PctCollegeGrad, TotalPop, CollegeGrad) %>%
  top_n(5, PctCollegeGrad) %>%
  arrange(desc(PctCollegeGrad))
#> # A tibble: 5 x 5
#>   NHGISCODE                 PLACE PctCollegeGrad TotalPop CollegeGrad
#>       <chr>                 <chr>          <dbl>    <int>       <int>
#> 1 G27071500         Woodland city       79.80769      312         249
#> 2 G27047104       North Oaks city       73.25481     3481        2550
#> 3 G27043270 Minnetonka Beach city       71.95402      435         313
#> 4 G27015616         Dellwood city       71.75989      733         526
#> 5 G27063544     Sunfish Lake city       65.74074      432         284
```

16) Which city had the highest increase in its proportion of college graduates?
Carver: +43.7

```
mn_change <- mn %>%
  filter(YEAR %in% c("1990", "2008-2012")) %>%
  group_by(NHGISCODE) %>%
  filter(n() == 2) %>% # Only places available for both years
  # Convert to negative for 1990 so we can add them
  mutate(PctCollegeGrad = ifelse(YEAR == "1990", -PctCollegeGrad,
PctCollegeGrad)) %>%
  summarize(PLACE = PLACE[1], ChangeCollegeGrad = sum(PctCollegeGrad))

mn_change %>%
  top_n(5, ChangeCollegeGrad) %>%
  arrange(desc(ChangeCollegeGrad))
#> # A tibble: 5 x 3
#>   NHGISCODE            PLACE ChangeCollegeGrad
#>       <chr>            <chr>             <dbl>
#> 1 G27010144      Carver city          43.68583
#> 2 G27000730 Albertville city          31.68633
#> 3 G27067036    Victoria city          30.12387
#> 4 G27025622  Greenfield city          29.53196
#> 5 G27026990     Hanover city          29.37340
```