IPUMS Data Training Exercise:

An introduction to IPUMS NHGIS

(Exercise 2 for R)

Learning goals

- Create and download a NHGIS data extract
- Decompress data file and read data into R
- Analyze the data using sample code

Summary

In this exercise, you will gain an understanding of how the NHGIS datasets are structured and how they can be leveraged to explore your research interests. Using NHGIS datasets, you will explore changes in the number of college graduates living in Minnesota cities to answer the research question: "Which cities in Minnesota saw the greatest change in the number of college-educated residents since 1990?"

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:

Code	Purpose
%>%	The pipe operator 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* ").
as_factor	Converts the value labels provided for IPUMS data into a factor variable for R
summarize	Summarize a dataset's 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
ggplot	Make graphs using ggplot2
weighted.mean	Get the weighted mean of the variable

Common Mistakes to Avoid

- Not changing the working directory to the folder where your data is stored.
- Mixing up = and ==; to assign a value in generating a variable, use "<-" (or "="). Use "==" to test for equality.

Registering with NHGIS

Go to https://www.nhgis.org/ click on Login at the top, and apply for access. On login screen, enter email address and password and submit! Then, enter the NHGIS data finder.

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 “Educational Attainment”, 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? ____

Find Data for the Period of Interest

With the topic already selected, click the Years filter button, then select “1990”, and submit the selection.

The Select Data grid now lists all the tables related to the topic of “Educational Attainment” with data from 1990. One way to proceed would be to select one of the “source tables” 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: “time series tables”, 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 “Persons 18 Years and Over by Educational Attainment [7]”? ________________________________

____

6. What type of Geographic Integration does this table use? _________________
7. In the Select Data grid, click on “Nominal” in the Geographic Integration column. With this type of integration, what should we keep in mind as we compare data across time? ____

____

____

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 structure.

- 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 “Place” geographic level. – (In census terminology, cities, villages, and town centers are all “places”)

Review and submit your extract

- Click the green VIEW CART button under your data cart.
- Review the 'Extract Request Summary' screen, describe your extract, and click Submit Extract.
- You will receive an email when the data is available to download.

- To access the page to download the data, follow the link in the email, or click on the Download or Revise Extracts link on the homepage.

Getting the data into your statistics software

The following instructions are for R. If you would like to use a different stats package, see: https://ipums.org/support/exercises

Download the data

- Go to https://data2.nhgis.org/main and click on Download or Revise Extracts.
- Right-click on the Data link next to the extract you created.
- Choose "Save Target As..." (or "Save Link As...").
- Save into "Documents" (Documents should pop up as the default location).
- Do the same thing for the DDI link next to the extract.
- (Optional) Do the same thing for the R script.
- You do not need to decompress the data to use it in R.

Install the ipumsr package

- Open R from the Start menu
- If you haven't already installed the ipumsr package, in the command prompt, type the following command:

```
install.packages("ipumsr")
```

Read the data

- Set your working directory to where you saved the data above by adapting the following command (Rstudio users can also use the "Project" feature to set the working directory. In the menubar, select File -> New Project -> Existing Directory and then navigate to the folder):

```
setwd("~/")
```

Note: "~/" goes to your Documents directory on most computers.

- Run the following command from the console, adapting it so it refers to the extract you just created (note the number may not be the same depending on how many extracts you have already made):

```
library(ipumsr)

ddi <- read_ipums_ddi("nhgis_00001.xml")

data <- read_ipums_micro(ddi)
```

Or, if you downloaded the R script, the following is equivalent: source("nhgis_00001.R")

- 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)

library(ggplot2)
```

- To stay consistent with the exercises for other statistical packages, this exercise does not spend much time on the helpers to allow for translation of the way IPUMS uses labelled values to the way base R does. You can learn more about these in the value-labels vignette in the R package. From R run command:

```
vignette("value-labels", package = "ipumsr")
```

- 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.

Part I: Analyze the Data

Part 1: 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 2008-2012, it will be helpful first to think about only the columns of interest. Defining "college graduates" as anyone with a bachelor's degree or higher, which columns should we use? *Note: The 2008-2012 data include both estimates and margins of error columns. For now, we'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 “CollegeGrad,” 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 “ChangeCollegeGrad”, and compute the total change in college grads between 1990 and 2008-2012 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 you would like to find out which cities had the greatest increases in the proportion of the population with bachelor'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 2008-2012? ___________________

```
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 2008-2012? _____________

Create a summary table with ChangePctCollege and calculate the differences between the PctCollege variables between 1990 and 2008-2012.

```
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 = "Educational Attainment"? 953

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

Find Data for the Period of Interest

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 5-Year

5. What advantage is there in using this table rather than the "Persons 18 Years and Over by Educational Attainment [7]"? A large portion of people aged 18-24 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 "geographic integration" 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 won't tell us how much of a city'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.

Part 1: Analyze the Data

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

9. Why do you think some places are missing values for certain years? Possibilities: They didn'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.

10. How many place records are there for Minnesota? 916

11. Defining "college graduates" as anyone with a bachelor's degree or higher, which columns should we highlight? "Bachelor's degree" for both years and the "Graduate or professional degree" for both years

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

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

Part II: College Grads by Place (Optional)

14. What was the total population 25+ of St. Paul in 2008-2012? 174,459

15. Which city had the highest percentage of college grads in 2008-2012? Woodland: 79.8%

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