IPUMS Data Training Exercise

An Introduction to IPUMS NHGIS

(Exercise 1 for R)

Learning goals

- Create and download an 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. This exercise will use an NHGIS dataset to explore slavery in the United States in 1830. You will create data extract and use a sample code to analyze these data.

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
weighted.mean	Get the weighted mean of the variable

Common Mistakes to Avoid

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

   Note: In this exercise, for simplicity we will use "weighted.mean". For analysis where variance estimates are needed, use the survey or srvyr package instead.

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!

Quick Instructions

Apply any combination of the four filters below to find 1830 slavery related tables:

- **Geographic Levels** = 'State'
- **Years** = '1830'
- **Topics** = 'Slavery'
- **Datasets** = '1830_cPop'

Guided Instructions

Suppose you were interested not only in slavery, but in all that's covered by the 1830 Census. **Note: To view all available 1830 data, use only the Years Filter set to '1830'.**

1. How many tables are available from the 1830 Census? _____
2. Other than slave status, what are some other topics could we learn about for 1830?

    _____

Locate the Desired Table

- Let's focus in on the slavery topic. To narrow the results, apply the Topics Filter of 'Slavery'. (You can find it at the bottom of the list of POPULATION topics.)
- The Select Data grid now lists all the tables related to the topic of Slavery. If you don't also have the Years Filter on, scroll down to find the 1830 tables, or utilize additional filters to further limit the available tables.
- Locate this 1830 table and answer the questions that follow: "NT12. Race/Slave Status by Sex"

Learn About the Table in the Data Finder

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

   _____

4. For which geographic levels is the table available?

   _____

5. Close the table pop-up window and inspect the Select Data table… What is the universe for this table? _____

   _____

   _____

6. What differentiates this table from the other available slavery tables from 1830?

   _____

   _____

7. Name a percentage or ratio this table would allow us to calculate that the other tables would not, based on the counts available in each table?

   _____

Make an 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 table NT12 to your Data Cart.
- (Optional) R is also capable of using shape files, if you want, you can download them by: – Click on the "GIS Boundary Files" Tab – Click on the plus sign to the left of the State Geographic Level Table
- Click the green Continue button in your Data Cart.
- On the Data Options screen, select the geographic level of "State".
- Click the green Continue button in 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), add an extract description if you wish, and click Submit.

Request and Extract the Data

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 "queued" to "in progress" to "complete", at which time you will be able to click the "tables" link to download the data.

- Return to the Extracts History page if not currently there.
- Right-click on the "tables" link for the extract you created.
- Choose 'Save Target As...' (or 'Save link as...').
- Save the zip file into "Documents".
- Repeat the process for the GIS data if you are going to use it(right-click, choose 'Save Target As...', ...)
- The R package can read the extracts as zip files, or if you wish to open in other programs, you can unzip them, by: Right-clicking on the 'nhgis0001_csv.zip' file, and select Extract All... Then click the Extract button. (Repeat for the shape if you desire).

Getting the data into your statistics software

The following instructions are for R.

Download the Data

- Go to https://data2.nhgis.org/main and click on Download/ Revise Extracts
- Right-click on the "data" link next to extract you created, under "files"

- Choose "Save Target As..." (or "Save Link As...")

- Save into "Documents" (that 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 #

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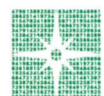
This tutorial will also rely on the dplyr package, so if you want to run the same code, run the following command (but if you know other ways better, feel free to use them):

```
library(dplyr)
```

Analyze the Data

Part 1: Analyze the Data

8. How many states/territories are included in this table?

   _____

```
length(table(nhgis$STATE))
```

9. Why do you think other states are missing? _____

   _____

```
table(nhgis$STATE)
```

10. Create a new variable called total_pop, with the total population for each state, by summing the counts in columns ABO001 to ABO006.
    Which state had the largest population? _____

```
nhgis <- nhgis %>%

    mutate(total_pop = ABO001 + ABO002 + ABO003 + ABO004 +

    ABO005 + ABO006)


nhgis %>%

    as.data.frame() %>%

    select(STATE, total_pop) %>%

    arrange(desc(total_pop)) %>%

    slice(1:5)
```

11. Create a variable called slave_pop, with the total slave population by summing the variables ABO003 and ABO004. Which state had the largest slave population?

    _____

EXERCISE 1 FOR R

```
 nhgis <- nhgis %>%

    mutate(slave_pop = ABO003 + ABO004)


nhgis %>%

    as.data.frame() %>%

    select(STATE, slave_pop) %>%

    arrange(desc(slave_pop)) %>%

    slice(1:5)
```

12. Create a variable called pct_slave with the Slave Population divided by the Total Population. Which states had the highest and lowest Percent Slave Population?

_____

```
nhgis <- nhgis %>%

    mutate(pct_slave = slave_pop / total_pop)


nhgis %>%

    as.data.frame() %>%

    select(STATE, pct_slave) %>%

    filter(pct_slave %in% c(min(pct_slave, na.rm = TRUE),
max(pct_slave, na.rm = TRUE)))
```

13. Are there any surprises, or is it as you expected? _____

_____

_____

```
nhgis %>%

    as.data.frame() %>%

    filter(pct_slave > 0.5) %>%
```

```
        select(STATE, slave_pop, total_pop, pct_slave)


nhgis %>%

    as.data.frame() %>%

    filter(STATE %in% c("New York", "New Jersey")) %>%

    select(STATE, slave_pop, total_pop, pct_slave)
```

Part II: Inspect the Codebook

Open the .txt codebook file that is in the same folder as the comma delimited file you have already analyzed. The codebook file is a valuable reference containing information about the table or tables you've downloaded.

Some of the information provided in the codebook can be read into R, using the function `read_ipums_codebook()`

14. What is the proper citation to provide when using NHGIS data in publications or researcher reports?

    _____

    _____

```
cat(ipums_file_info(nhgis_ddi, "conditions"))
```

15. What is the email address for NHGIS to share any research you have published? (You can also send questions you may have about the site. We're happy to help!)

    _____

Part III: Make maps using R (Bonus)

One of the reasons we are excited about bringing IPUMS data to R is the GIS capabilities available for free in R. To use them, you'll need to install the sf package with the following command:

```
install.packages("sf")
```

If that doesn't work, or you prefer the older style "sp" package for geographic analysis, ipumsr does provide support. For more information, see the "ipums-geography" vignette in R.

To load the NHGIS data with the spatial features attached, we use this command (again, you may need to adjust the filepaths):

```
# Change these filepaths to the filepaths of your downloaded
extract

nhgis_csv_file <- "nhgis0001_csv.zip"

nhgis_shp_file <- "nhgis0001_shape.zip"
```

16) Make a map of the percent of the population that are slaves.

```
nhgis <- read_nhgis_sf(

data_file = nhgis_csv_file,

shape_file = nhgis_shp_file,

verbose = FALSE

)

# Calculate percent enslaved again
nhgis <- nhgis %>%
    mutate(
    total_pop = AB0001 + AB0002 + AB0003 + AB0004 + AB0005 +
    AB0006,
    slave_pop = AB0003 + AB0004,
    pct_slave = slave_pop / total_pop
)


# Note the function `geom_sf()` is a very new function, so you
may need to update ggplot2 to run.

library(ggplot2)
```

EXERCISE 1 FOR R

```
if ("geom_sf" %in% getNamespaceExports("ggplot2")) {

    ggplot(data = nhgis, aes(fill = pct_slave)) +

        geom_sf() +

        scale_fill_continuous("", labels = scales::percent) +

        labs(

            title = "Percent of Population that was Enslaved by

            State",

            subtitle = "1830 Census",

            caption = paste0("Source: ", ipums_file_info(nhgis_ddi,
"ipums_project"))

        )

}
```

Answers

1. How many tables are available from the 1830 Census? <u>Fifteen (15)</u>

2. Other than slave status, what other topics of interest could we learn about for 1830? <u>Population that is urban, particular ages, deaf and dumb, blind, and foreign born not naturalized.</u>

Learn About the Table in the Data Finder

3. How many variables does this table contain? <u>Six (6)</u>

4. For which geographic levels is the table available? <u>Nation, State, & County</u>

5. What is the universe for this table? <u>Persons</u>

6. What differentiates this table from the other available slavery tables from 1830? <u>It includes the counts of "white" persons, in addition to "colored" persons</u>

7. Name a percentage or ratio this table would allow us to calculate that the other tables would not, based on the counts available in each table: <u>Percentage of total population in slavery, or ratio of slave: free population</u>

Part 1: Analyze the Data

8. How many states/territories are included in this table? <u>Twenty-Eight (28)</u>

9. Why do you think other states are missing? <u>In 1830, there were not any other states yet! Every decennial census is a historical snapshot, and NHGIS provides census counts just as they were originally reported without "filling in" any information for newer areas.</u>

```
#>
#> Alabama              Arkansas Territory       Connecticut
#> 1                    1                        1
#> Delaware             District Of Columbia     Florida Territory
#> 1                    1                        1
#> Georgia              Illinois                 Indiana
#> 1                    1                        1
```

```
#> Kentucky          Louisiana              Maine
#> 1                 1                      1
#> Maryland          Massachusetts          Michigan Territory
#> 1                 1                      1
#> Mississippi       Missouri               New Hampshire
#> 1                 1                      1
#> New Jersey        New York               North Carolina
#> 1                 1                      1
#> Ohio              Pennsylvania           Rhode Island
#> 1                 1                      1
#> South Carolina    Tennessee              Vermont
#> 1                 1                      1
#> Virginia
#> 1
```

10. Create a new variable called total_pop, with the total population for each state, by summing the counts in columns ABO001 to ABO006. Which state had the largest population? <u>New York</u>

```
#> # A tibble: 5 x 2
#> STATE total_pop
#>
#> 1 New York 1913006
 #> 2 Pennsylvania 1348233
#> 3 Virginia 1211405
#> 4 Ohio 937903
#> 5 North Carolina 737987
```

11. Create a variable called slave_pop, with the total slave population by summing the variables ABO003 and ABO004. Which state had the largest slave population?

    <u>Virginia</u>

12. Create a variable called pct_slave with the Slave Population divided by the Total Population. Which states had the highest and lowest Percent Slave Population?

    <u>South Carolina (54.27%) and Vermont (0.00%)</u>

13. Are there any surprises, or is it as you expected? <u>Possibilities: Did you know some states had more slaves than free persons? Did you know that some 'free states' were home to substantial numbers of slaves?</u>

Part II: Inspect the Codebook

14. What is the proper citation to provide when using NHGIS data in publications or researcher reports?

    <u>Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 13.0 [Database]. Minneapolis: University of Minnesota. 2018.</u> http://doi.org/10.18128/D050.V13.0 <u>(Check for most current citation here: https://nhgis.org/research/citation)</u>

15. What is the email address for NHGIS to share any research you have published? (You can also send questions you may have about the site. We're happy to help!)

    <u>nhgis@umn.edu</u>

Part III: Make maps using R (Bonus)

16. Make a map of the percent of the population that are slaves.

```
# Change these filepaths to the filepaths of your downloaded
extract

nhgis_csv_file <- "nhgis0001_csv.zip"

nhgis_shp_file <- "nhgis0001_shape.zip"
```

```
nhgis <- read_nhgis_sf(

     data_file = nhgis_csv_file,

shape_file = nhgis_shp_file,

verbose = FALSE )

 # Calculate percent enslaved again

nhgis <- nhgis %>%

     mutate(

          total_pop = ABO001 + ABO002 + ABO003 + ABO004 + ABO005
          + ABO006,

          slave_pop = ABO003 + ABO004,

          pct_slave = slave_pop / total_pop

   )


# Note the function `geom_sf()` is a very new function, so you
may need to update ggplot2 to run.

library(ggplot2)

if ("geom_sf" %in% getNamespaceExports("ggplot2")) {
     ggplot(data = nhgis, aes(fill = pct_slave)) +

          geom_sf() +

          scale_fill_continuous("", labels = scales::percent) +
          labs(

               title = "Percent of Population that was Enslaved
               by State",

               subtitle = "1830 Census",

               caption = paste0("Source: ", ipums_file_info
               (nhgis_ddi, "ipums_project"))

     )

}
```

EXERCISE 1 FOR R