A User-Friendly U.S. Census Browser for R

Kiegan Rice

1 Introduction

Census data is an important snapshot of information about a country at different times throughout their history. It is an integral part of keeping track of the history of a country and the people who occupy it, as it gives us records of the occupants of a country and how they live their lives. The value of census data to a country and its citizens is difficult to overstate. Many recognize the usefulness of the census data, especially when aggregated and presented in a way that allows for viewers to learn something new about the world around them. While the history books we learn from in high school present us with a narrative of the events that occured, students often don't get to interact with the raw data ourselves in that learning environment. Accessible and usable census data allows the exploration of different demographic groups over time, or investigations of a particular period of time and what the demographic and economic landscape looked like in the past. Even today, for those who can't travel the world, data about the world around them could allow them to learn more about those places and learn something more about groups of people they may not usually engage with.

From an early point in the United States' history, there were "many eminent men of science" who recognized the value of the census data and worked to aggregate and present the data that had been collected on the population. Francis A. Walker's "Statistical Atlas of the United States", based on the 1870 census, was an impressive effort in aggregating population data to present it in a visually appealing way. Although the Census Bureau's "Statistical Atlases" eventually stopped being made, they were an important start to the effort of presenting census data to a wider public.

Today, as methods of data analysis and visualization continue to be developed and improved, access to census data allows us to look back on that history and explore, synthesize, and visualize the information to learn more about patterns in many different parts of the population.

Mention Dr. Hofmann's paper to improve the Statistical Atlas here... forthcoming? published? As well as Haley's ggmosaic stuff

Ever-improving visualization and data-wrangling methods in R give those interested in statistical graphics a wealth of abilities to explore and learn from data; however, it is difficult to make use of these tools on census data if that data is not available and easy to explore in one location.

An inherent problem in census data is that a country's census changes over time; the variables collected, how they are collected, and even the locations they are collected for change as the country is formed, and subsequently grows and changes. The United States census data is no exception to this rule. In a little under two and a half centuries, the census has taken on many different forms. Data on occupations has transformed as the employment landscape has changed; new states have been formed, the most recent being within the last 100 years; definitions of various demographic groups and the terminology used to describe them have been updated as the demographic makeup of the country has changed. Each decennial census brings a different set of variables to the table. Sometimes these variables are new things the Census Bureau is interested in learning, while sometimes they remove variables that are no longer relevant or whose information is captured somewhere else.

Unfortunately, because the founders of the U.S. Census were unable see 200 years into the future, those interested in working with census data are left with quite the inescapable mess. If you want to focus in on a particular demographic group and their journey as part of the population of the United States, you may have ten or more different variables names to describe that one group over the course of the census from 1790 to 1960 - and that is just for one single group! Of course, we cannot just simply change variable names to match our own research needs. It is important to keep the data in its true form and be honest to the way that the population was defined at different times throughout history, even if our instinct may be to 'clean' the data by changing variable names.

This, of course, leaves the user with a wide variety of variables that are far from consistent across years. In order to track one demographic group across years - let alone many groups - a clean user interface that helps users see exactly what information is available to them across years is a necessity. To streamline the process and assist researchers in finding out what information they have access to and what they don't, we present a U.S. Census Browser for R, with the user interface being a Shiny application, and the downloadable files being 'tidy' csv files that those with a small amount of R experience should be able to work with.

2 Background ?(Literature Review)?

The University of Virginia Library hosted a "Historical Census Browser" for many years that allowed users to search United States Decennial Census Data for use in research, teaching, and personal inquiry. The data included data on various aspects of the U.S. population from the 1790 Census through the 1960 Census, originally populated using the ICPSR 3 dataset. This Historical Census Browser was free and available for use for anyone with an internet connection. This Census Browser allowed a user to peruse available topics for each census year, at both the state and county levels. The Historical Census Browser was taken down on December 31, 2016, with the county-level aggregated data becoming unavailable several months before this. The source was widely used, with several other university libraries and educational resources including University of California Santa Barbara, University of Pennsylvania, University of Michigan(University of Michigan Population Studies Center 2017), and the Smithsonian's History Explorer directing researchers to use the University of Virginia source.

The University of Virginia Libraries website now directs users to "Social Explorer" or the "National Historical Geographic Information System" (NHGIS) website. Social Explorer, which is populated with the ICPSR 2896 data, requires that users pay to use (or use through library access from a library who has paid for it), and does not offer full download. NHGIS, hosted through the University of Minnesota, is very difficult for users to navigate when looking for specific information across multiple years. The Institute for Social Research at the University of Michigan (ICPSR), has the full ICPSR 2896 data set, split into 106 separate data sets, but requires that users be a part of a member institution, and requires users to agree that they will not distribute the information in any way after gaining access to it. It also does not have any browser function for users to look at the data unless they download the entire file for each year, in either a SAS, SPSS, ASCII, or Stata set-up.

None of the aforementioned resources provide the full advantages that the Historical Census Browser offered: a free-to-use and user-friendly data browser that allows users to choose which data they are interested in using, and download the complete (aggregated) records for their own independent use. Although the Historical Census Browser data is now dated based on the existence of an updated,

cross-checked version of the decennial census through ICPSR (2896), it is for the most part accurate and allows for free and open use for researchers and others alike. *This is a bad sentence, figure out a better way to state this.*

3 Work

3.1 Data

Although the county-level data had already been removed from the website, the state-level aggregated records for each decennial census from 1790 through 1960 were captured from the website in September of 2016 and saved as raw data with the intent to create a resource for R users that allowed the same main functionalities of the original University of Virginia Historical Census Browser, streamlined for easy data searches and data management.

The majority of files for the decennial censuses (sp?) were complete upon capture from the website. However, there were two years of census data that are missing some column names and thus are unverified data. For the 1890 and 1920 censuses (sp?), columns were compared to all columns from ICPSR 02896 and ICPSR 3 data files by both correlation and Euclidean distance, and thus some columns were able to be correctly identified.

Describe this process in more detail once you actually get this shit figured out. How many columns were correctly matched? How many weren't? Were we able to at least get a foothold on which columns go where in the excel spreadsheet after identifying a couple of them?

Each year of the dataset began in a separate file, with rows being each State, each column being a demographic variable, and values being state-aggregated counts of number of people in that category. Each state was given a label `Type` as either a State or Territory, as some U.S. territories participated in the decennial census before they received full statehood. Each individual file can be found as `states` followed by the year of the census, (e.g. `states1790`) in this package. Each of these datasets were combined into one list, `stateslist`, which is the list of data that is used to populate the census browser.

3.2 Description

This package includes the state-aggregated data for each individual year, as well as a list of all of the years together. The intended interaction with these data sources is via the Shiny application, `Get Your Data`. This interactive Shiny application presents the user with two options - focusing on a single year of the census, found in the `Single Year` tab, or focusing on multiple years at once, found in the `Multiple Years` tab.

The `Single Year` tab allows users to choose which year they are interested in from a drop-down menu, as well as up to two variables of interest they want to search for within the data for that year. For example, if a user is interested in the prevalence of farms in 1860, they can select 1860 and then search for "farms" in the first "Variable of Interest" spot (see Figure). Users can then see a complete list of all variable names in that particular year that have "farms" in their title. With this list now available, users can select each variable they want by clicking on it. Once users have selected all variables they want to download data for, clicking on the "Download Filtered Data" button will open a window in which the user can choose a file name and directory location for saving

the resulting csv file. Variables will not remain selected if the user chooses to change their search term or add a second search term. However, if the user has interest in many separate variables that require different search terms, it is quite easy to combine multiple csv files in R once they have been downloaded. *Should I write a function for this?? To be an exported function within the package? That seems like a good idea...* The variables `STATE`, `YEAR`, `TOTAL.POPULATION`, and `TYPE` will always be included in the resulting csv file, whether the user selects them or not.

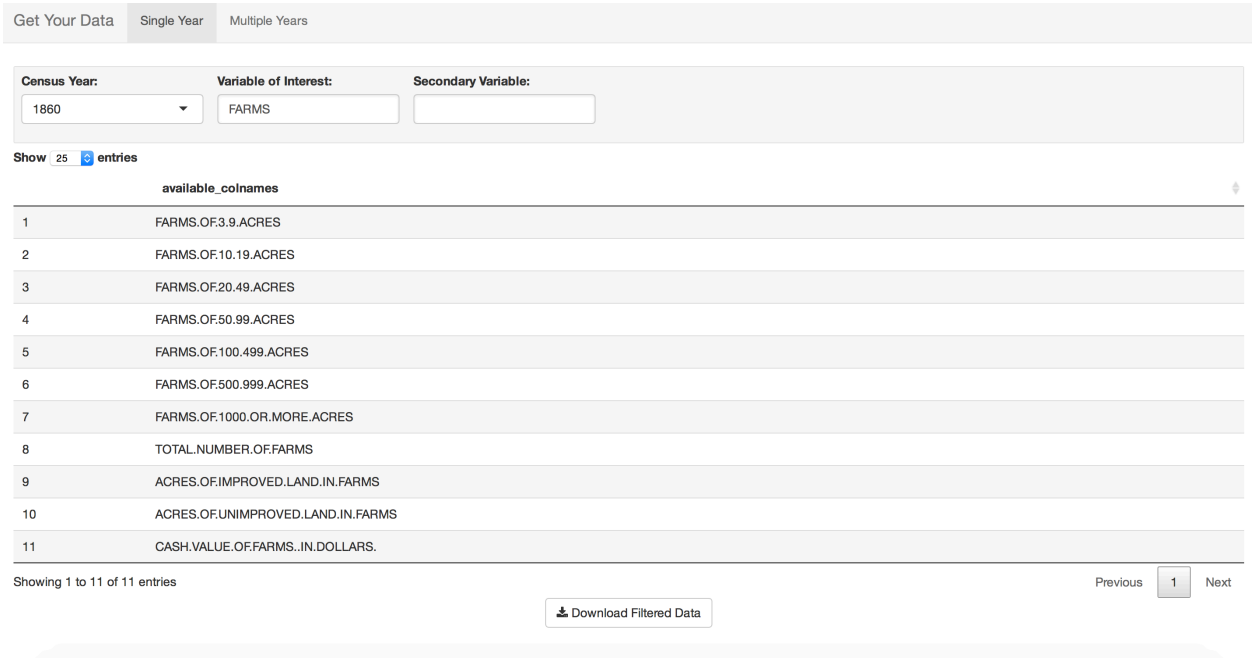

Figure 1: The resulting view of a search for "farms" in the 1860 U.S. Census on the Single Year tab of the Get Your Data app.

The `Multiple Years` tab allows users to look for variables over a range of years. It offers all decennial censuses (sp?) from 1790 to 1960, and offers the same search tool as before, but only allows users to narrow down results by one search term rather than the two provided in the `Single Year` tab. Users can select the range of years they are interested in by using the slider tool to specify their range of interest, and subsequently use the "Variable of Interest" search bar to narrow down their results. The results, being two-dimensional rather than just a list of available variables, are presented somewhat differently in the `Multiple Years` tab. Each available variable is listed along the left side of the interface, and each selected year is presented across the top as a column in the table. For each variable in the table, an `X` indicates for which years that particular variable is

present. Far to the right, there is a column which denotes a count how many of the selected years have that particular variable. Users can choose to order the results by most years present to least years present, if they choose. This is where it is important to keep in mind that variable names for different demographic groups change drastically over the course of the decennial census, and thus users will want to be cognizant of varying search terms that may need to be checked. *Should I still try to add a feature that pops up when they look for certain search terms? This seems hard to do without singling out a few groups, since pretty much every group has changed terminologies over time.*

Similar to the output file of the `Single Year` tab, once a user selects all variables they are interested in, they can select the "Download Filtered Data" button to download a csv file of the chosen variables for all states across the selected years. For years in the selected range that don't have a particular variable, the csv file will be filled in with NA values, which allows the structure of the data table to remain intact although there are some year-variable combinations that do not exist in the data. As mentioned previously, because of the changing nature of variable names in the decennial census, users may have to search for several separate terms, download each file separately, and combine them once the files are downloaded. However, they are fairly easy to combine because the structure ensures that key variables needed to add on new columns are in place, regardless of which variables users choose to download. *Should I also write a function to combine different data tables from different year ranges / different variables?* This results in all information about all variables of interest across the years of interest being together in one structured data table. Users can easily filter on specific variables or years while still maintaining the original full data set, which is advantageous for visualizing how a demographic group changes over time.

3.3 Example

We can easily use the "Get Your Data" Shiny app to search for a topic of interest, find the data we want, download state-level information, and tell a visual story of populations in the United States over time. To demonstrate the utility of this Shiny app, we will walk through an example on the history of the African American population in the United States. We begin by running the Shiny app:

```
library(shiny)
shiny::runGitHub("kiegan/censusbrowseR", subdir = "shiny")
```

The default set of years is 1790 to 1920. However, we can easily expand the slider range to be able to explore data across all the available years.

The available variable names here are already ordered by how many years they appear in, so we can see right off the bat that we have a serious lack of variable continuity. The identifying variables that appear each year are `YEAR`, `STATE`, `TOTAL.POPULATION`, and `TYPE`. That means the total number of times they appear is 10, the number of census years we are looking at. After that, the next most common variable only appears in 4 of the 10 years we have chosen. That means we are up against some pretty significant hurdles to track any demographic group over time.

For now, we are focusing in on African Americans throughout U.S. history, which means we need to start with the term `SLAVE` in the early years of the U.S. census. The vast majority of African Americans were slaves when the United States was founded, so variables that count the number of slaves are important variables to have in order to get a grasp on the African American population

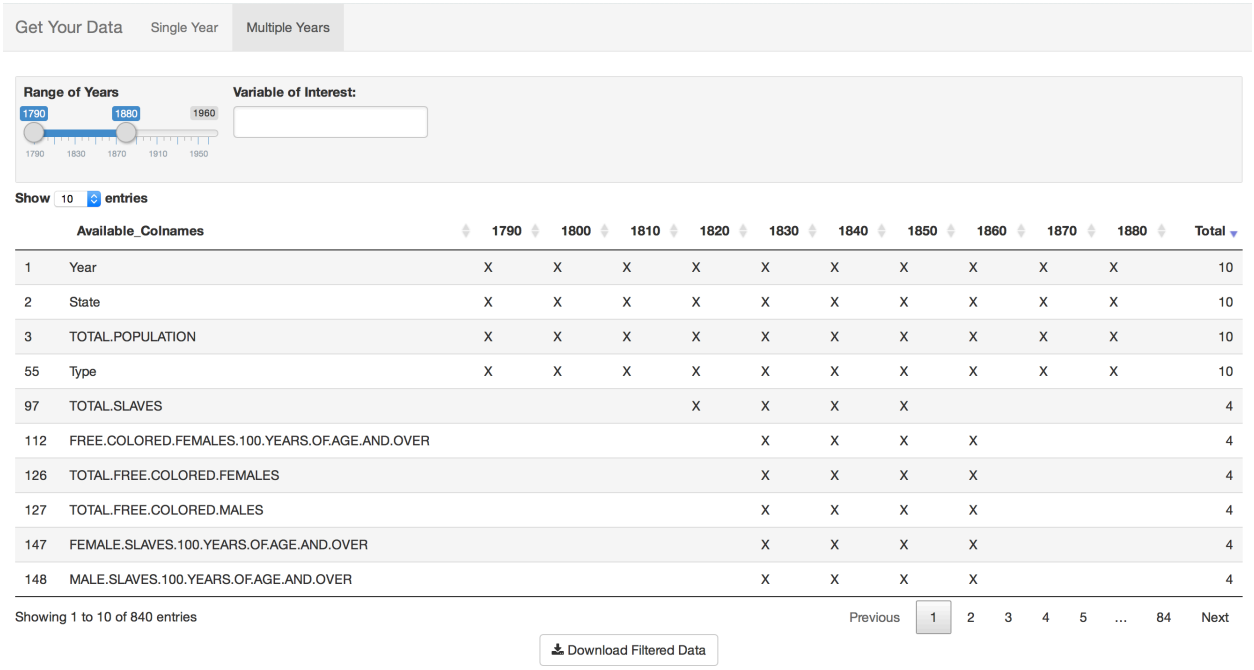

Figure 2: An example of available column names for the years 1790 to 1880.

in early U.S. history. It is important to note that the `SLAVE` categorization is the only term the U.S. census had related to African Americans in the early years of the census, so it is also the only source of information we have for that time period.

Once we have our range of years chosen, we simply enter a search term. Searching for `SLAVE` gives us a list of all the possible variable names across all years, and we can again sort by how how many years each variable name appears in. The results of this search can be seen in Figure ___.

For this example, I executed three separate searches within the `Get Your Data` app, using three different terms that the census used to categorize African Americans at different points in the decennial census. This gave me three separate .csv files, which I was able to easily combine using `full_join` from the `dplyr` package. Now, for an entire demographic group, all of the state-aggregated population counts are together in one data frame, although different terms are used at different points in time.

We can now begin to explore this data. For any particular year, we can subset the total data set to include only information for that year, so that we can more quickly determine which variables we have available for that particular year. For example, after subsetting on the year 1790, we can very easily plot state-level counts of the `SLAVES` variable for that year. Use of the `USAboundaries` package in concert with this census browser is recommended, as it provides the most accurate boundaries of the United States at any given date. Just as variable names and the make-up of the census rapidly changed throughout the United States' history, the boundaries of the states and territories changed as well. The recorded state-aggregated values in the dataset are for the states as they were during that census year, and thus sometimes include demographic counts from areas that differ from the current defined boundaries.

We now have a visualization of the slave population in the United States in 1790, and we can easily look at subsequent years of the decennial census to see how the population changes over time.

If we jump to the year 1820, and again filter on that particular year, we learn new information about the African American population.The census now includes a count of `TOTAL.FREE.COLORED.PERSONS`, as now there are citizens who are African American or other minorities that are free, and not slaves. At this time, many minorities were grouped together in the "colored persons" category, and as before, we do not have any other record for African Americans except this categorization and the "slaves" categorization. However, since the 1820 census included records for both `TOTAL.FREE.COLORED.PERSONS` and `TOTAL.SLAVES`, we can also now visualize the percentage of African Americans in each state that were categorized as "free persons". Knowing that both of these variables are available allows us to make a powerful comparison and visualize the growing divide in the United States at that time.

Moving forward from there, we can glean even more information about this population. After subsetting on the year 1850, we see that although there is still a `TOTAL.SLAVES` column in the record, many states did not record this variable. This is due to the abolition of slavery and these states being "free states" already. We can transform these `NA` values into zeros to account for this difference in the data and allow us to still calculate the percentages of free African Americans. Starting in 1850, we can also start investigating what the total African American population looks like in each state, rather than just the slave population and free population. This is particularly interesting as slavery is abolished and some migration begins to occur out of certain areas. The balance of the total African American population in each state begins to change somewhat as this post-slavery period begins.

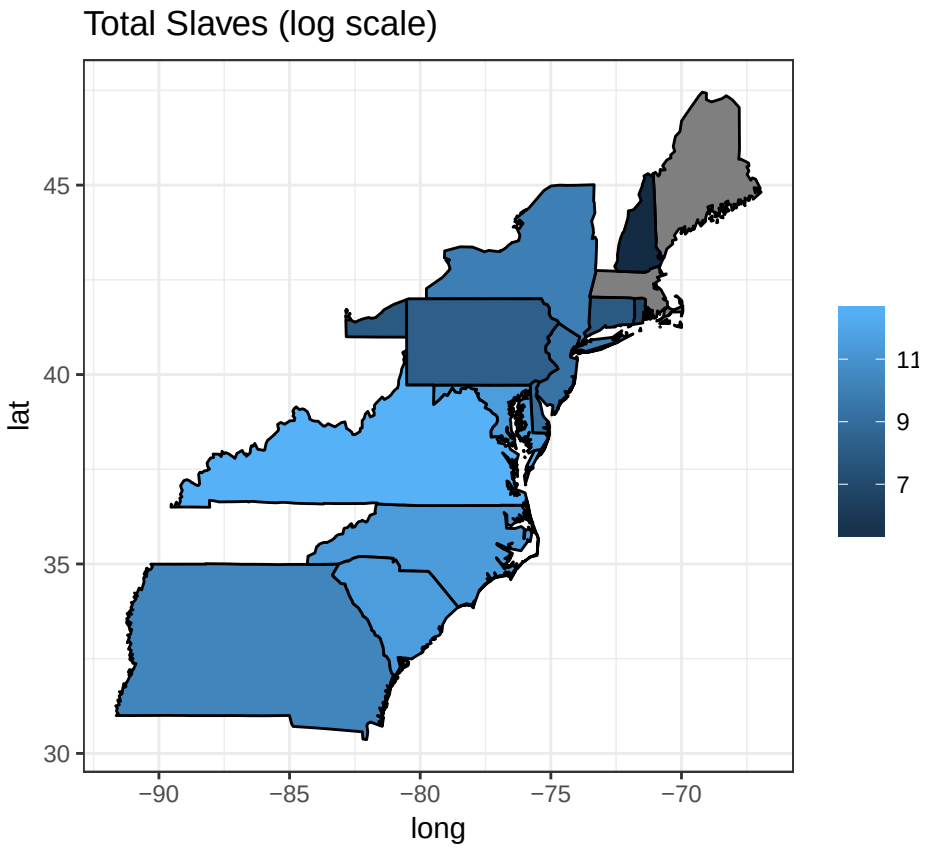

Figure 3: Total number of slaves per state in 1790, plotted on a continuous log scale. State boundaries for July 4, 1790 were gathered from `USAboundaries` package.

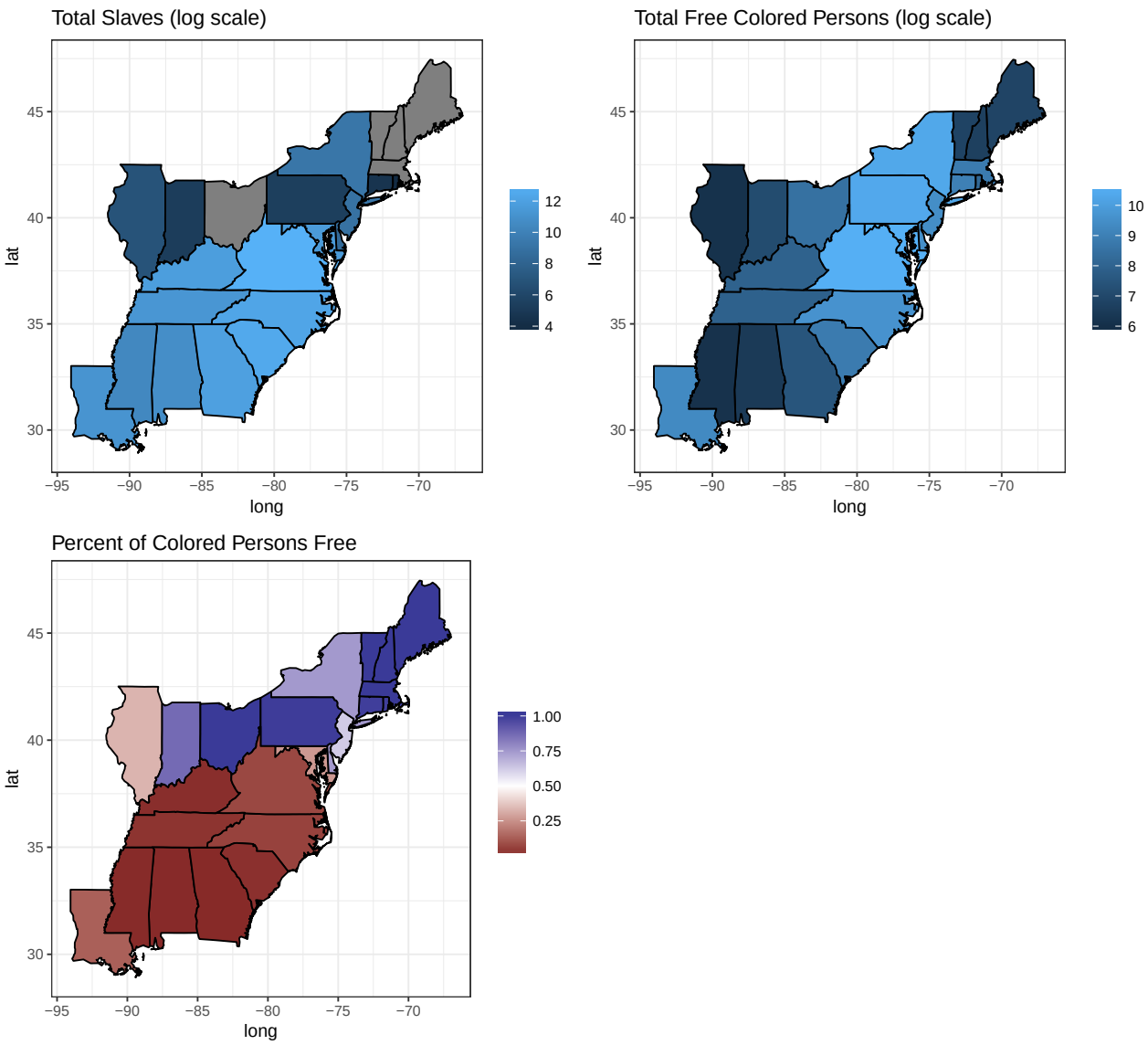

Figure 4: Number of Slaves, Number of Free Colored Persons, and Percentage of Free Colored Persons in 1820.

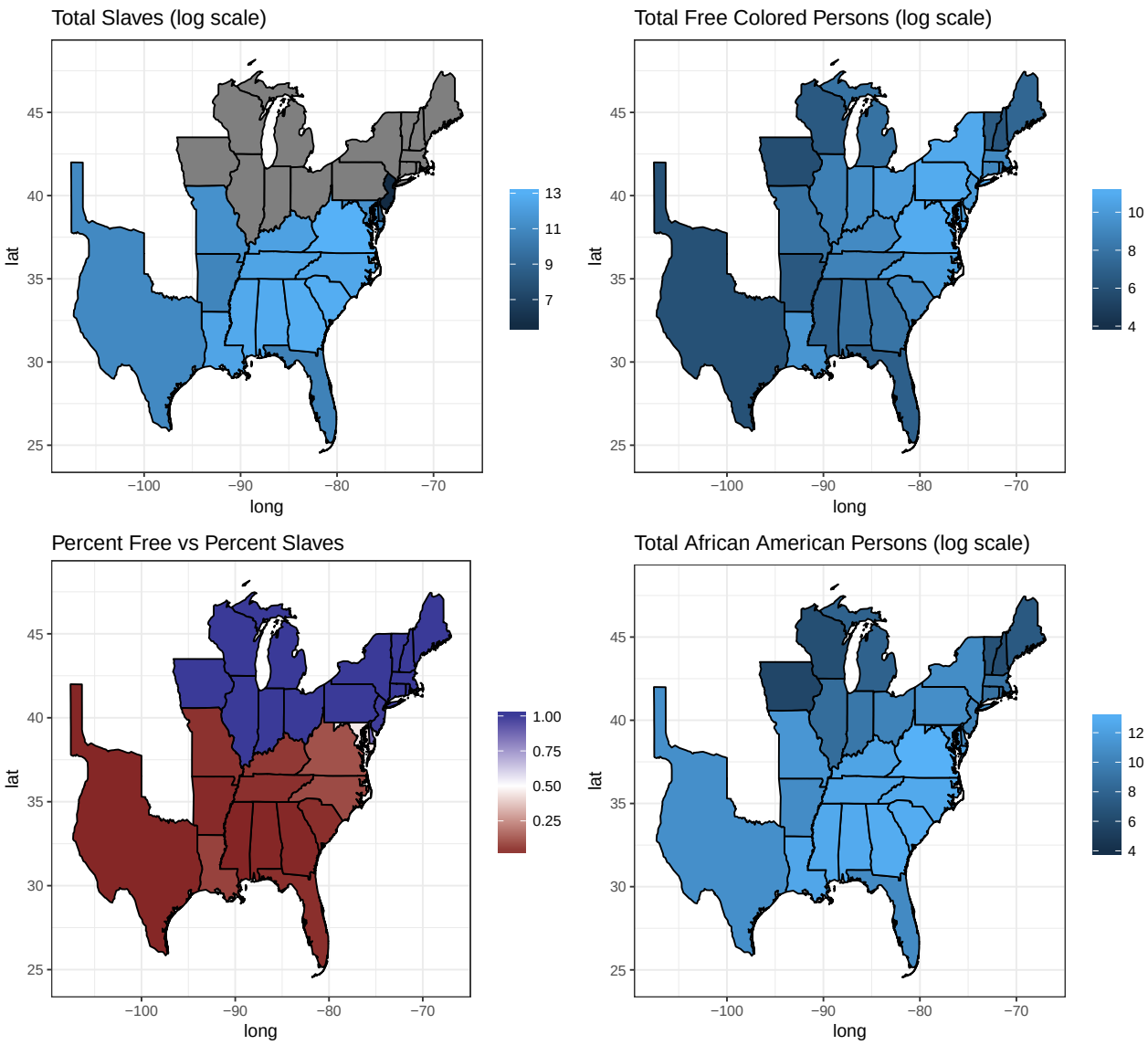

Figure 5: Number of Slaves, Number of Free Colored Persons, Percentage of Free Colored Persons, and Total African American Persons in 1850.

3.4 Implementation

How to use it - this is the most technical part

4 Discussion of Future Work

- County-level data would be great
- Current data (up to 2010 Census)

5 References

University of Michigan Population Studies Center. 2017. "Historical Census Browser (1790 - 1960)." Accessed March 27. http://www.psc.isr.umich.edu/dis/data/resource/detail/1369.