Accessing Data with the Census Bureau API

Alex Shum

March 25, 2014

1 Introduction

The United States Census Bureau has been conducting a decennial census since 1790. Originally this census was a simply to count the population across the country. More recently the decennial census includes a short-form asking for name, sex, age, and a few other demographic variables. About one in six households also received a long-form that contained additional socioeconomic questions. After the 2000 decennial census many of the long-form questions were collected as part of a new survey: the American Community Survey (ACS).

The ACS is an ongoing yearly survey that collects additional demographic variables including but not limited to age, sex, race, income and education. Unlike the decennial census, the American Community Survey is distributed based on a random selection of addresses every year. Although the ACS is only sent to a sample of all US households, this data is meant to provide more up to date information than the Census Bureau's decennial census. Both the decennial census and the American Community Survey are required to be completed by law; however it should be noted that the Census Bureau has not opted to prosecute anyone for failure to complete the decennial census or the ACS. Despite the lack of enforcement, the ACS still reports a response rate of 97%.

Both the decennial survey and the ACS data are used in part by federal, state and local agencies to allocate state funding and for policy decisions. The Census Bureau has also released some of this data for public use. Many of the data sets are available directly in a compressed format from the Census Bureau's FTP site: `http://ftp2.census.gov/`. Since 2012, the Census Bureau has also included an online developer's API in order to improve accessibility of the ACS and decennial census datasets. The Census Bureau's online API can be accessed online: `http://api.census.gov`.

We will discuss how the ACS data is structured when we request data and how to access data from the Census Bureau's online developer's API. We will also discuss what kind of variables are available and some limitations with the API. We will base this discussion on a paper by Stangl, Rundel, and Morgan [1] as a starting point on some of the limitations of the API. This article explores some multivariate frequency distributions using data from the ACS dataset; however, there are some gaps in what we can access and inconsistencies in the database.

2 Requesting Data

To access data from the census bureau online API we need to construct a proper HTTP GET request. A valid GET request is formed through a constructed web URL and we specify which dataset, year, variable and geographies that we are requesting. The basic structure of an HTTP GET request for the decennial census and for the ACS is as follows:

```
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=[GEOGRAPHY]
```

[DEVELOPER'S KEY] is an id code required to perform a valid GET request. A developer's key uniquely identifies everyone who requests data from the API. Requesting a key can be done by registering at

`http://www.census.gov/developers/tos/key_request.html`.

[YEAR] and [DATASET] specify the dataset and year of the data requested. The available datasets include the decennial census and the ACS. The ACS datasets are available in 1-year, 3-year and 5-year timeframes. The [YEAR] variable for the ACS datasets indicates the final year in the timeframe. For example, the 2012 5-year ACS dataset this is the ACS dataset that spans 2008-2012 and the 2012 3-year ACS dataset is the ACS dataset that spans 2010-2012. For the decennial census the [YEAR] indicates the year the census data was collected. When we request a [DATASET] we use abbreviations for the dataset we want: the ACS 5-year dataset is *acs*5. See table 1 to see which timeframes are available for each dataset and the abbreviations.

DATASETS	YEAR	abbr
Decennial Census	1990, 2000, 2010	sf1
ACS 5-year	2010, 2011, 2012	acs5
ACS 3-year	2011, 2012	acs3
ACS 1-year	2011, 2012	acs1

Table 1: Datasets and Years

If we wanted to request some data from the 2010 decennial census we would format our HTTP GET request as follows:

```
http://api.census.gov/data/2010/sf1?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=[GEOGRAPHY]
```

Similarly requesting data from the 2011 ACS 3-year dataset would require the following HTTP GET request:

```
http://api.census.gov/data/2011/acs3?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=[GEOGRAPHY]
```

[GEOGRAPHY] describes the geographic region of interest. We can choose varying levels of geographic areas including the entire United States:

```
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=us:*
```

We can also specify states using varying levels of detail:

```
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=state:*
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=state:06
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=state:01,06
```

The above HTTP GET requests specify all states, a specific state (California), or multiple states (Alabama and California) respectively. Some geographic regions are nested within larger regions and we can specify the containing region. For example, states contain counties and we can form an HTTP GET request for a certain county within a specific state:

```
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=county:*&in=state
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=county:037&in=sta
```

In the above HTTP GET requests we specify all counties in California and Los Angeles County in California respectively. Finally, there are even smaller geographic regions that we can specify multiple containing regions. For example, census tracts are within counties and states. We can specify a certain census tract within a county:

```
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=tract:*&in=state:
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=tract:101110&in=s
```

In the above HTTP GET requests we specify all census tracts within Los Angeles County and census tract 1011.10 within Los Angeles County respectively.

Specifying a correct geographic region or combination of regions and specifying available demographic variables ([VARIABLES]) require a more detailed knowledge of how the census datasets are organized. We discuss geographies in section 3 and variables in section 4. We also discuss how the data is formatted and structured in section 4.

3 Geography

The Census Bureau has a very sophisticated system of hierarchy for geographic entities. For the ACS, at the top level there is the entire nation, followed by region, division, state, county, county-subdivision, tract, block group, place, congressional district, zip code area, school district and a few other geographic divisions. See table 2 for a complete table of geographic entities available on the census API for the 2012 ACS.

Summary Level	Description
010	us
020	region
030	division
040	state
050	state-county
060	state-county-county subdivision
140	state-county-tract
150	state-county-tract-block group
160	state-place
250	american indian area/alaska native area/hawaiian home land
310	metropolitan statistical area/micropolitan statistical area
320	state-metropolitan statistical area/micropolitan statistical area
330	combined statistical area
340	state-combined statistical area
350	new england city and town area
400	urban area
500	state-congressional district
510	state-congressional district-county
610	state-state legislative district (upper chamber)
620	state-state legislative district (lower chamber)
795	state-public use microdata area
950	state-school district (elementary)
960	state-school district (secondary)
970	state-school district (unified)

Table 2: List of valid geographic combinations for 2012 ACS 5-year

From table 2, there is a specific hierarchy of geographic regions and specific valid combinations of geographic regions. Different ACS datasets might have slightly different geographic regions available. For example, in the 2010 decennial census if our geographic area is zip code tabulation areas then we are required to specify states. By contrast, the 2012 ACS 5-year dataset has zip code tabulation areas but we do not need to specify states. The 2010 ACS 5-year dataset simply does not have zip code tabulation areas available. Fortunately, each dataset available on the Census Bureau online API include an associated geography file formatted in JSON and a similar file formatted in XML. JSON stands for JavaScript Object Notation and it is a lightweight machine format for sending data. JSON is structured using name-value pairs and is also designed to be human-readable. There are many libraries to generate and process JSON. The JSON formatted file for geographies has the following format:

```
{
  "name": "tract",
  "requires": [
    "state",
    "county"
  ],
  "optionalWithWCFor": "county"
}
```

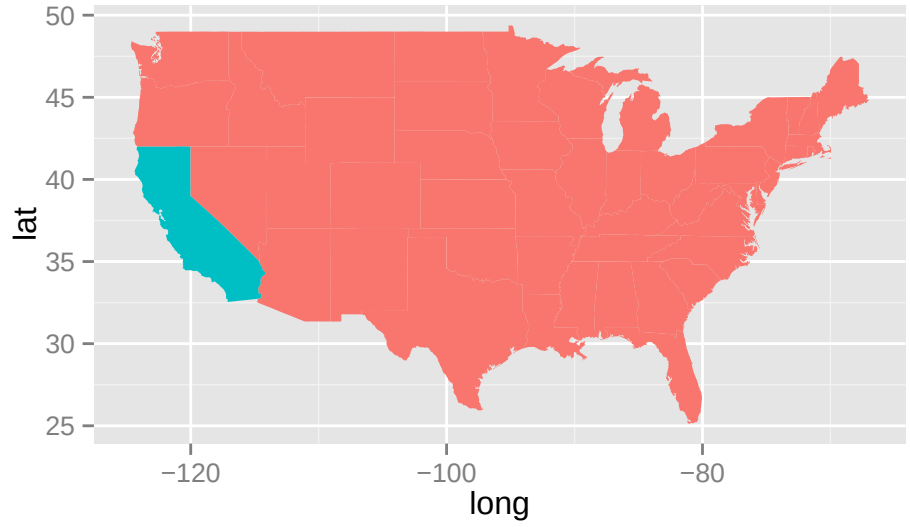


Figure 1: California Selected

This JSON file specifies that for census tract level geography we are always required to specify a state. We are required to specify a county in certain cases. If we want all census tracts within a state then we are required to specify state. If we want all census tracts within a county then we need to specify a state and a county. If we want a specific census tract, we must specify both a county and a state. This is due to how census tracts are labelled; it's possible that census tracts located in different states and counties have the same label.

The same geographic information is also available in XML. XML stands for Extensible Markup Language and is similar in structure to the HTML format used for webpages. XML is another format for sending and storing data. It is formatted in a tree-like structure with a hierarchy of categories with associated values. Here is the same geographic area information formatted in XML:

```
<fips name="tract">
  <requires name="state"/>
  <requires name="county" is-optional-with-wcfor="true"/>
</fips>
```

In the following examples, we will examine various geographies available from the online API and see which combinations of geographic regions are required. We will use the United States as a whole and various geographic areas within the state of California for our examples.

At the top of the geography hierarchy we can form valid HTTP GET requests specifying a country-wide geography; this is data aggregated among all states and corresponds to summary level 010 from table 2. Below that we specify a state-level geography. At this level we can view data for all states or select a particular state; this is summary level 040. In figure 1 we have selected California with the following HTTP GET request: `http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'SKEY]&get=[VARIABLES]&for=state:06`.

Below states we can select counties and census tracts. At the county level, after we specify California we can look at counties within California (summary level 050). We can see from figure 2 that we have selected Los Angeles county within California state. This corresponds to an HTTP GET request of the following form: `http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'SKEY]&get=[VARIABLES]&for=county:037&in=state:06`. There are a few other geographic regions we can subset by within a state such as ZIP code tabulation area; for example we can select 90210 which corresponds to Beverly Hills,

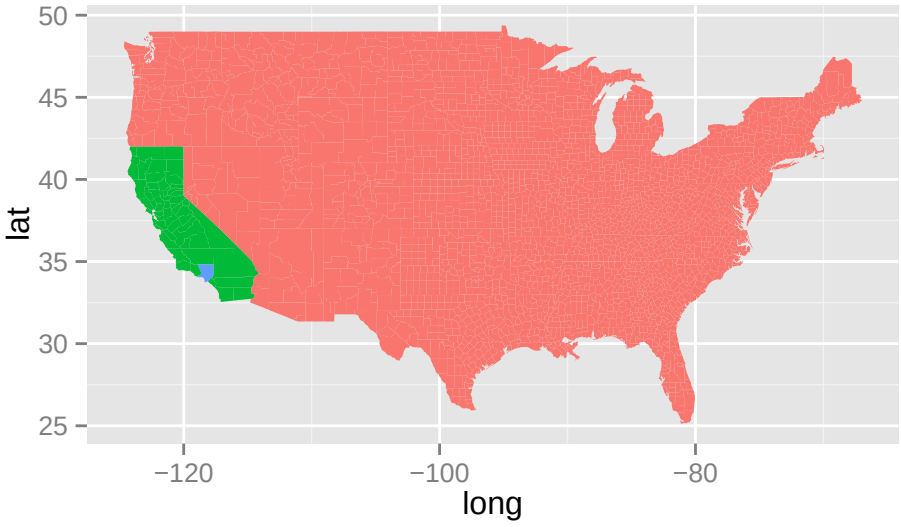


Figure 2: Los Angeles county selected

California.

There are a number of valid geographic entities below the state and county level. For example, there are school districts, county subdivision, metropolitan statistical areas and legislative districts. In order to form a valid HTTP GET request, smaller geographic divisions often require us to specify the containing state or county. In figure 3 we have selected Pasadena, a county subdivision, within Los Angeles County. For county subdivisions we cannot simply specify Pasadena and we also cannot specify just California and Pasadena; subdivision alone and state-subdivision are not a valid geographic combinations. Instead, we must specify California, Los Angeles County and then Pasadena. We can see from table 2 that state-county-subdivision is a valid summary level.

Although all the entries in table 2 are valid geographic combinations, some parts of geographic combinations may be optional. For example, if we wanted census tracts we need to specify state and county (summary level 140). In contrast to county subdivisions where we needed to specify both state and county, for census tracts specifying county is optional. This means that state-tract is also a valid geographic combination. It is best to refer to each datasets JSON or XML file for geographic compatabilities.

Due to the sheer number of different geographic entities there are geographic entities which cannot be used together for valid HTTP GET requests. Some of the smaller geographic divisions are not necessarily nested in one of the larger geographic divisions; for example, ZIP code areas are generally used by the United States Postal Service and might span different counties or census tracts. Additionally, Legislative districts do not line up with county borders and school districts often do not line up with either legislative districts or county borders. ¡EXAMPLE HERE¿

For a more detailed look at which geographies need to be specified, refer to the census bureau list of summary levels for each dataset or the XML or JSON geography files. For the 2012 ACS dataset this is located at `http://api.census.gov/data/2012/acs5/geo.html`, the XML file is located at `http://api.census.gov/data/2012/acs5/geography.xml` and the JSON file is located at `http://api.census.gov/data/2012/acs5/geography.json`.

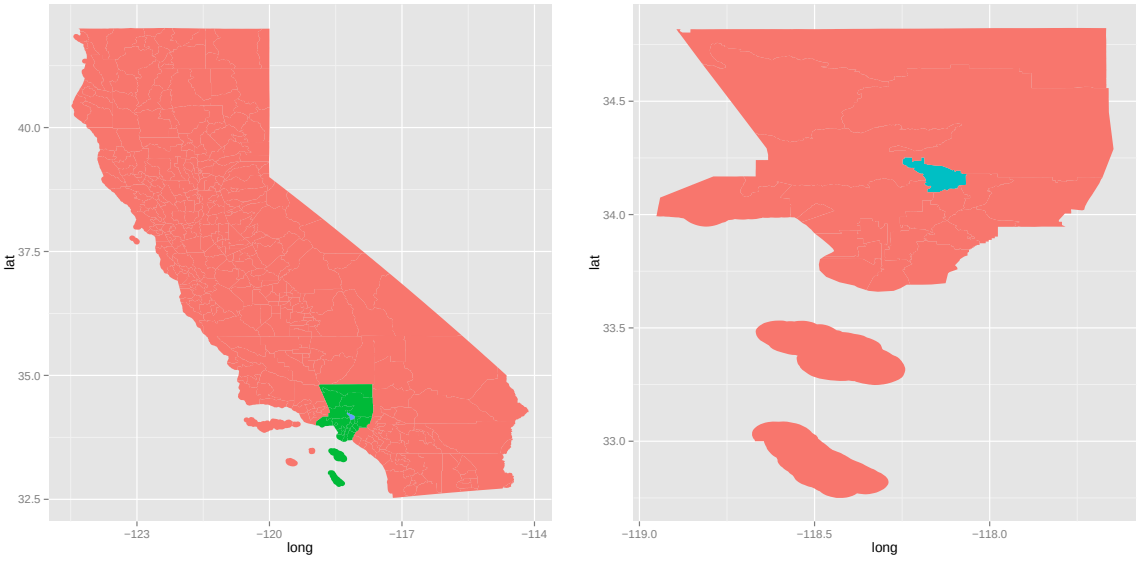


Figure 3: Pasadena county subdivision selected

4 Finding Data Sets and Tables Structure

The Census Bureau API includes a master index of available dataset formatted in both JSON and XML. The JSON file is available online at `http://api.census.gov/data.json` and the XML file is available online at `http://api.census.gov/data.xml`. This master index includes necessary meta-information about each dataset including description, links to geography and variable information and contact information for maintainer of datasets. The JSON format is formatted as follows:


```
{
  "c_vintage": 2012,
  "c_dataset":[
    "acs5"
  ],
  "c_geographyLink": "http://api.census.gov/data/2012/acs5/geography.json",
  "c_variablesLink": "http://api.census.gov/data/2012/acs5/variables.json",
  "c_tagsLink": "http://api.census.gov/data/2012/acs5/tags.json",
  "c_examplesLink": "http://api.census.gov/data/2012/acs5/examples.json",
  "c_documentationLink": "http://www.census.gov/developers/",
  "c_isAggregate": true,
  "title": "2012 American Community Survey: 5-Year Estimates",
  "webService": "http://api.census.gov/data/2012/acs5",
  "accessLevel": "public",
  "bureauCode": [
    "006:07"
  ],
  "contactPoint": "Census Bureau Call Center",
  "description": "The American Community Survey (ACS) is a nationwide survey...",
  "identifier": "2012acs5",
  "mbox": "pio@census.gov",
  "publisher": "US Census Bureau",
  "references": [
    "http://www.census.gov/developers/"
  ],
```

```
  "spatial": "US",
  "temporal": "2012"
  },
}
```

The above excerpt from `http://api.census.gov/data.json` is the meta-information about the 2008-2012 ACS 5-year dataset. For this dataset there are links to the associated geography file and to another JSON file *variables.json* which is a list of variables available from this dataset. For the above example the same meta-information is formatted in XML as follows:

```
<dataset vintage="2012"
  geographyLink="http://api.census.gov/data/2012/acs5/geography.xml"
  variablesLink="http://api.census.gov/data/2012/acs5/variables.xml"
  tagsLink="http://api.census.gov/data/2012/acs5/tags.xml"
  examplesLink="http://api.census.gov/data/2012/acs5/examples.xml"
  documentationLink="http://www.census.gov/developers/"
  pod:webService="http://api.census.gov/data/2012/acs5" isAggregate="true"
  pod:accessLevel="public"
  dcat:contactPoint="Census Bureau Call Center"
  dct:identifier="2012acs5"
  pod:mbox="pio@census.gov"
  dct:publisher="US Census Bureau"
  dct:spatial="US"
  dct:temporal="2012">
<dataset-name> <part name="acs5"/> </dataset-name>
<dct:title>2012 American Community Survey: 5-Year Estimates</dct:title>
<pod:bureauCode> 006:07 </pod:bureauCode>
<dct:description>
The American Community Survey (ACS) is a nationwide survey...
</dct:description>
<pod:reference link="http://www.census.gov/developers/"/>
</dataset>
```

After the user has selected the dataset of interest, they will need to lookup what variables are available for that dataset. The JSON and XML master index of datasets contain the location of *variables.json* and *variables.xml* respectively which list available variables available for that dataset. Available variables are organized into tables referred to as "concepts"; a "concept" is a combination of factors. For example, "Health Insurance Coverage Status by Sex by Age" is a concept from the 2012 ACS. Although health insurance coverage, sex and age are all different factors, this "concept" contains information that links these three factors together.

Within each table there are sub-tables that provide information on different levels of each of the factors. For example, under the "Health Insurance Coverage Status by Sex by Age" concept there is a table for males under 6 with health insurance. For this concept there are tables for each combination of levels for sex (male, female), age group (under 6, 6 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 to 74 and 75 and over) and health insurance coverage (with and without health insurance). Additionally there are a number of total columns: total number of males, total number of females and totals for males/females in each age group.

To lookup what variables are available, the *variables.json* and *variables.xml* files contain a list of all sub-tables along with a description. The JSON formatted *variables.json* is formatted as follows:

```
"B27001_056E": {
  "label": "Female:!!75 years and over:!!With health insurance coverage",
  "concept": "B27001.  Health Insurance Coverage Status by Sex by Age"
},
```

This describes sub-table 056E of concept B27001. This is a table of 75 and older females with health insurance. The associated XML formatted version is formatted as follows:

```
<var xml:id="B27001_056E"
  label="Female:!!75 years and over:!!With health insurance coverage"
  concept="B27001. Health Insurance Coverage Status by Sex by Age"/>
```

5 Limitations

The dataset used by Stangl, Rundel, and Morgan [1] is a random subset of the 2010 ACS public use microdata sample. This article contains a number of classroom exercises that ask the reader to calculate some basic proportions about various demographic data. We will attempt to use the Census Bureau online API to examine the same demographic variables.

Using the Census Bureau's online API for these exercises presents us three main problems. The first problem is the structure of how the data is organized in the public use microdata sample versus how the data is organized in the online API. The second problem is that the census does not provide proportions and finding standard errors for these proportions requires a bit more work. Finally, the third problem is that certain combination of variables are simply not available from the online API.

	Sex	Age	Married	Income	HoursWk	Race	USCitizen	HealthInsurance	Language
1	0	31	0	60.00	40	white	1	1	1
2	1	31	0	0.36	12	black	1	1	0
3	1	75	0	0.00		white	1	1	0
4	0	80	0	0.00		white	1	1	0
5	1	64	1	0.00		white	1	1	0
6	1	14	0			white	1	1	0

Table 3: Random subset of 2010 ACS public use microdata sample

Data from the ACS public use microdata sample is organized to describe individuals. Each row of the dataset describes an anonymized individual and each column represents a different demographic variable. See table 3 for a small subset of the data used by Morgan et al.

	Total	M Total	M<6	M<6 w/insurance	M<6 w/o insurance
Alabama	4693822	2256713	186155	176591	9564
Alaska	686905	349855	33053	29026	4027
Arizona	6304406	3099407	279297	249163	30134
Arkansas	2862023	1394466	120828	115053	5775
California	36783532	18138870	1561623	1461559	100064
Colorado	4949633	2457605	210948	193227	17721

Table 4: Health Insurance Coverage information from Census API. Each of the M columns indicate male, subsequent numbers indicate age; there are corresponding F columns for females that is not shown to conserve space.

By contrast the online API does not provide individual specific data. When we perform a HTTP GET request we must specify what geographic level of detail we want. The geographic level of detail does not go below the county subdivision or census tract level. Instead of individual specific data we have data that has been aggregated for an entire geographic region. This is likely due to privacy reasons; if we have individual specific data for a dozen demographic variables along with geographic information it might be possible to reveal this individual's identity.

In table 4 we specify state level summaries for health insurance coverage status for various age groups by gender (from the census API this is table B27001). Each variable is in the form of a table is referred to as a *concept* and each column in table 4 is referred to as a *label*. The first label of each concept is an overall total subsequent labels are subsets of this overall total. We've conveniently renamed column headers in table 4; the naming convention from data provided by the API is to enumerate each with the table name $B27001\_001$, $B27001\_002$ etc.

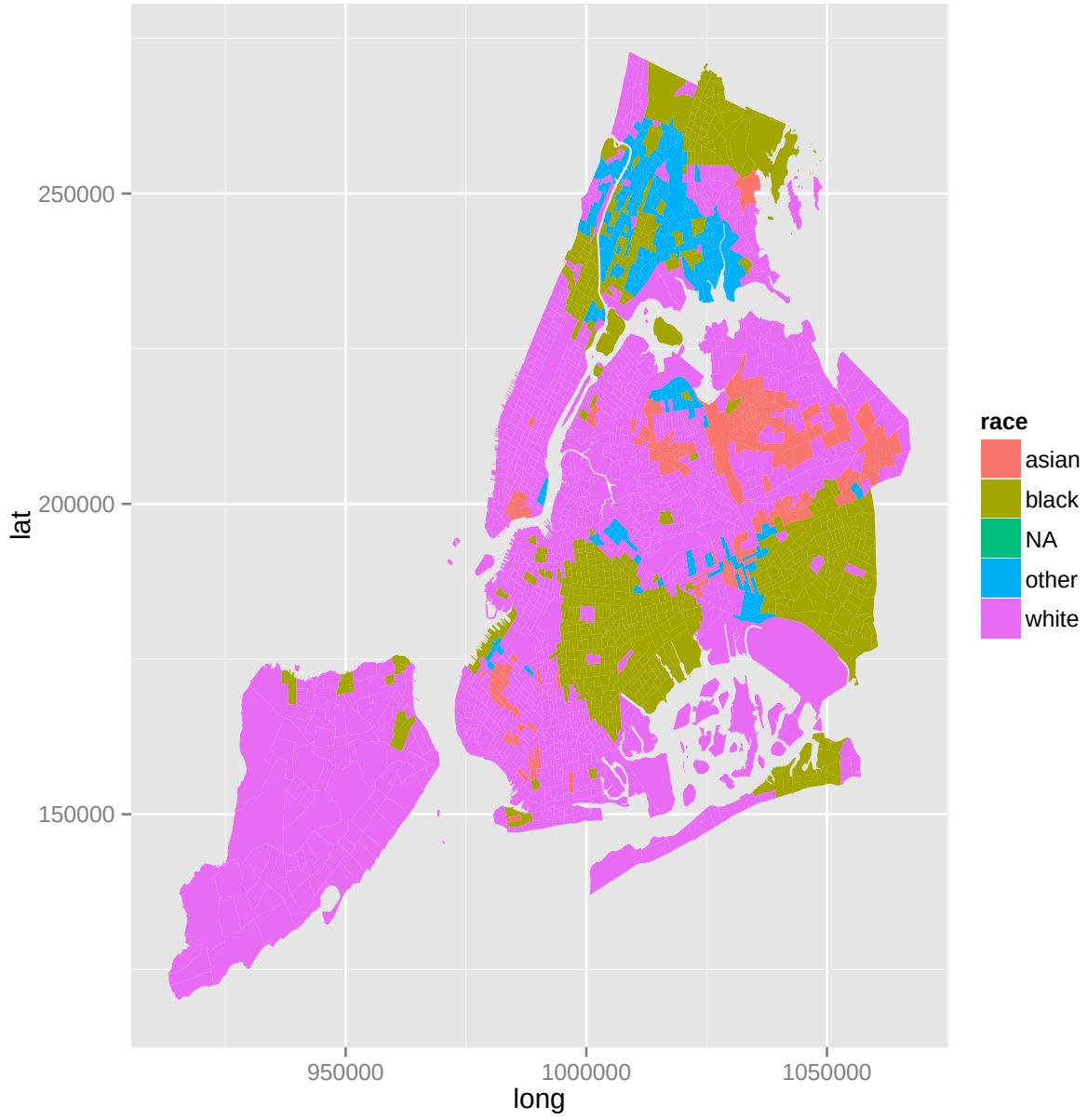
The way this data is organized is not tidy <include hadley tidy data reference>. Instead it appears that the columns contain additional categorical information: gender, age and insurance. In table 5 we have tidied up the data so that each row is an observation and each column is a variable. Originally table B27001 contains a number of *total* columns: overall total, total number of males, total number of females, and within each gender a total number of people within an age group. In reshaping this data, we felt that these total columns were redundant once the data is in a tidy form.

	state	gender	age	insurance	freq
1	Alabama	m	<6	yes	176591
2	Alaska	m	<6	yes	29026
3	Arizona	m	<6	yes	249163
4	Arkansas	m	<6	yes	115053
5	California	m	<6	yes	1461559
6	Colorado	m	<6	yes	193227

Table 5: Reshaped Health Insurance Coverage data.

6 Other Issues

7 Examples from ACS



8 Conclusion

References

[1] Dalene Stangl, Mine Çetinkaya Rundel, and Kari Lock Morgan. "Taking a Chance in the Classroom: The American Community Survey". In: *CHANCE* 26.1 (2013), pp. 42–46. DOI: 10.1080/09332480.2013.772392. eprint: http://www.tandfonline.com/doi/pdf/10.1080/09332480.2013.772392. URL: http://www.tandfonline.com/doi/abs/10.1080/09332480.2013.772392.