Lecture 7: Data Visualization Exercises

SDGB-7844

October 24, 2018

American Community Survey Data

The American Community Survey (ACS) is an ongoing survey by the U.S. Census Bureau. It regularly gathers information previously contained only in the long form of the decennial census, such as ancestry, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics. These data are used by many public-sector, private-sector, and not-for-profit stakeholders to allocate funding, track shifting demographics, plan for emergencies, and learn about local communities. Sent to approximately 295,000 addresses monthly (or 3.5 million per year), it is the largest household survey that the Census Bureau administers.

The Census Bureau selects a random sample of addresses to be included in the ACS. Each address has about a 1-in-480 chance of being selected in a given month, and no address should be selected more than once every five years. Data is collected by internet, mail, telephone interviews and in-person interviews. Approximately one third of those who do not respond to the survey by mail or telephone are randomly selected for in-person interviews. About 95 percent of households across all response modes ultimately respond.

The Census Bureau aggregates individual ACS responses (i.e. microdata) into estimates at many geographic summary levels. Among these summary levels are legal and administrative entities such as states, counties, cities, and congressional districts, as well as statistical entities such as metropolitan statistical areas, tracts, block groups, and census designated places. Estimates for census blocks are not available from ACS.

ACS estimates are available via a number of online data tools.American Fact Finder (AFF) is the primary tool for disseminating ACS data, allowing users to drill down to specific tables and geographies (starting with 2013 estimates, AFF also includes block group data). A selection of the most popular tables are shown in QuickFacts. Other tools include OnTheMap for Emergency Management, Census Business Builder and My Congressional District. My Tribal Area featuring 5-year estimates for federally recognized tribes, launched in 2017. The Summary File is the most detailed data source, and is available as a series of downloadable text files or through an application programming interface (API) for software developers. Kyle Walker's *tidycensus* R package will be our main source of data for the following exercises.

Household Income

During our study of Airbnb listing prices in previous lectures, we discovered that there was a moderate difference in the average rental price per night in New York City. In this exercise, we will use data from the American Community Survey to compare our Airbnb listing data to household income.

1. What variables are available to measure household income from the American Community Survey (2016)?
2. How does the American Community Survey define household income? Is this important information to report under assumptions for your analysis?
3. Why might we want to use a measure of centrality such as a median rather than the mean?
4. Does the American Community Survey provide the margin of error for their estimates of measures of household income?

Exploratory Data Analysis

5. Using the *tidycensus* API, download the list of variables available for the 2016 American Community Survey.

6. At what level (State, County, City) is the American Community Survey data available? How is this different than the deccenial census?
7. What is the variable name for median household income?
8. Use the *tidycensus* API to assign the median household income data for New York County (at the census tract level) to a data frame.
9. What are the margin of errors for this estimate? Add that measure to your data frame if you have not done it already.
10. Visualize your estimates for median household income and each tract's margin of error using ggplot2.
11. What does this plot tell us about the variation of income across census tracts in New York County? Is it easy to identify where these census tracts are in New York County?

Geo-spatial Analysis and Visualization

12. Using the *leaflet* mapping library for R (or another mapping library of your choice), visualize the estimates for median household income by census tract.
13. Display the following data in the "tooltip" when mousing over your plot: Median Household Income Estimate, Margin of error (+/- estimate) and Census Tract Name.
14. Using the Airbnb listings data, visualize data for the top 100 listings in the Airbnb data set. Explain which variable you will use to rank the top listings in the data set.
15. What options does the *leaflet* package give us for visualizing points on a map? Choose the visualization that you feel is most appropriate for mapping multiple listings. Map the 500 most expensive listings in the data set.
16. Create a new data frame using the *tidycensus* API with data on median household income from Kings County, NY (Brooklyn). Join this data together with the data from New York County. Use ggplot2 to visualize median income for each county on the same plot (Hint: try facet wrap!)