CS&SS/STAT 563 — Statistical Demography — Spring 2020 - Homework no. 7

Due Monday May 25 at 2:00pm on the course Canvas website.

Note: These questions are linked—the overarching goal is to build a small dataset combining digital and administrative data. Read the whole assignment through to the end before starting, but start collecting data for question 1 early.

Note: Much of the geographic part of this assignment would be easier with a Google Maps API key, but I won't ask you to sign up for one. Even though the Google Maps API allows a certain number of free requests, it still requires billing information, which isn't appropriate for a homework assignment!

1. **Social media data.** *Goal: collect geolocated tweets from Twitter.*

   (a) Choose a US state or metropolitan area, and collect tweets from Twitter's streaming API using a geographic filter. You may have to look up bounding box information, either with `rtweet::lookup_coords()` or online. Try to stream tweets for as long as you can—a few hours, or a full day.

   (b) Report the number of tweets you collected. Look at a few of them yourself, and produce some sort of visualization of the distribution of tweets. This could be a time series plot, or a map if you're ambitious.

   If you're unable to collect tweets: If Twitter authentication doesn't work for you, sign up for a different API instead and collect data from it using `httr`. Ideally, pick an API with geographic data. One option is the Yelp API, which provides data on restaurants and other business. You can adapt demo code from here: https://github.com/ccgilroy/yelp-restaurants

2. **Census data.** *Goal: collect ACS or Census data to compare to the Twitter data.*

   (a) Read this tutorial on accessing the US Census data API with `tidycensus`: https://csde-uw.github.io/tidycensus-tutorial/. Install the `tidycensus` and `tigris` packages, and sign up for a Census API key.

   (b) Use the same geography from (1), and with `get_acs()` get the total population (variable `B01001_001`) for an appropriate subgeography. If you chose a state, maybe this is counties. If a metro, maybe it's tracts or zctas (zip codes).

   (c) Plot the data you collected somehow. (Again, maps are great if you're ambitious! The `sf` package pairs nicely with `ggplot2`.)

3. **Combine data sources.** *Goal: aggregate tweet counts and link them to a Census geography.*

   Note: This is a challenging question. Give it a try, but don't be concerned if you can't get all the steps to work.

   (a) Extract latitudes and longitudes from the Twitter data with `rtweet::lat_lng()`. Think about whether using all geolocation information (the default) makes sense here.

   (b) Identify which Census geography those points are located in, based on the subgeography you chose in (2). There are multiple ways to do this, but `tigris::append_geoid()` may work for you (note: rename `lng` to `lon`). You may need to geocode only a small sample of your data, depending on its size.

   (c) Group tweets into counts by subgeography, and join to the Census data. (`tidyverse` tools work well for this.)

   (d) Fit a simple statistical model using total population to predict tweet counts (or vice-versa). Display and discuss your results.