# Notes for coding map of national pharmacy deserts

Great summary of tidycensus in general: <https://walker-data.com/census-r/index.html>

Summary of spatial data in R in general: <https://cengel.github.io/R-spatial/intro.html>

## 1. Geocode all pharmacy locations

* Read in pharmacy address data, clean addresses
  + Compare to WA pharmacies to see any differences
* Divide pharmacies into states
* Run geocode in batches until I hit $200 each month
* Pricing strategy: ([link to Google information](https://developers.google.com/maps/documentation/geocoding/usage-and-billing))
  + $5 per 1000. If I want to hit $200 that means I can geocode 40,000 per month. So do over 2 months.
  + $200 free per month is automatically applied.
  + Instructions for [how to set up Google Cloud](https://developers.google.com/maps/documentation/geocoding/cloud-setup) account
    - Dashboard for my [pharmacy-deserts-analysis](https://console.cloud.google.com/home/dashboard?project=pharmacy-deserts-analysis) project
* How to

## 2. Read in shape files for all states

* Use function get\_decennial(year = 2020).
* Geography = loop function for all of US states, then counties, then tracts

## 3. Variables of census for low income and population

* Two good sites for looking up variable names for specific topics:
  + Variable table for API: <https://api.census.gov/data/2021/acs/acs5/variables.html>
  + Data explorer census.gov <https://data.census.gov/table?q=Race+and+Ethnicity&tid=ACSDT1Y2021.B02001>
  + DECENNIAL TABLES: PAGE 139 in this: <https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/summary-file/sf1.pdf>
    - Note that decennial is best for getting more accurate population estimates. But, it doesn’t have enough social variables (e.g., income, vehicle, etc.) to define pharmacy deserts let alone the characteristics of pops living in them.
    - Consider- define pharmacy desert census tracts with ACS variable estimates. Then get total population in that tract using decennial data as more accurate estimate?
* Could try R function “load variables” as outlined on tidycensus blog here ([link](https://walker-data.com/tidycensus/articles/basic-usage.html)), then search table in R to find the names of variables I want… seems tedious
  + But could look up the exact name I want from ACS and search that? Age!! Etc.
* load\_variables(year = 2010, dataset = "sf1") %>%
* view()
* Demographic and Housing Characteristics summary files, which will include other variables typically included in the decennial Census data (age & sex breakdowns, detailed race & ethnicity), expected available in May 2023 according to data.census.gov
  + Good resource for decennial notes: <https://walker-data.com/umich-workshop-2022/intro-2020-census/#1>
  + Action: for now, keep the ACS variables in the code. If I ever get to this coding stage, I can check what is available in decennial tables at that point in time.

**Using the margin of error estimate appropriately:**

* Helpful slides with basic overview of what margin of error vs conf interval vs standard error is and how theyre calculated: <https://www.sacog.org/sites/main/files/file-attachments/3-understanding_moes_and_cvs_in_acs_sacog.pdf?1445383556>
* Textbook. <https://ccrpc.org/wp-content/uploads/2015/02/american-community-survey-guide.pdf> . Page 3: “Similar to the aggregated count data, margin of error must be calculated when using sample data for proportions, ratios, products, and percent change. The following equations should be used to calculate the margin of error when calculating estimates.”
* Ran Wei 2023 great article on MOE in census as applied to residential segregation including good background on census sampling sizes: <https://link.springer.com/article/10.1007/s11113-023-09754-6>
* Boscoe 2022 another great article: <https://www.sciencedirect.com/science/article/pii/S235282732200057X>
* “A compass for understanding and using ACS Data: What Researchers need ot know.” Census 2009. <https://www.census.gov/content/dam/Census/library/publications/2009/acs/ACSResearch.pdf>
* “Best Practices for Reporting American Community Survey in Municipal Planning” <https://www.ccrpcvt.org/wp-content/uploads/2018/10/ACS_Guide_Final_20181003.pdf> .
  + MOE for this population is higher than the estimate itself so data is not reliable, and the data should not be used. ACTION: In code, add checks here to drop any columns for which MOE is > the estimate.
    - And/or. (Boscoe 2023): We removed census tracts with fewer than 30 households due to measurement unreliability. This value corresponds roughly to a minimum population size of 100 that has been used previously (Diez Roux et al., 2001) given an average household size of about 3 (United States Census Bureau, 2021).
  + This plan uses the coefficient of variation (CV) for data points to determine whether margins of error are too high for a data point to be reliable. When the CV is below 15%, data are considered highly reliable, and these estimates are presented without caveat. When the CV is between 15% and 30%, the plan indicates that the data should be used with caution. Data with a CV over 30% are not reported to avoid unreliability. Methodology drawn from Jurjevich et al. “Navigating Statistical Uncertainty.” Journal of the American Planning Association 84, no. 2 (Spring 2018): 112-126. Year-to-year changes between ACS estimates are statistically different unless otherwise reported
* Advanced users like the Wei article will use the Variance Replicate Tables. (links in this helpful doc: <https://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/MultiyearACSAccuracyofData2015.pdf>)
  + 2021 VRT information and links: <https://www.census.gov/programs-surveys/acs/data/variance-tables.html>
  + VRT tables available for download here: <https://www2.census.gov/programs-surveys/acs/replicate_estimates/2021/data/5-year/>
  + “This leads to an exact variance and MOE for user-derived measures by taking into account the covariance between ACS estimates.”
* See Starsnic downloaded slides!! “Margins of Error the ACS Way: Working with Variance Replicate Estimate Tables”
* Notes for R and MOE
  + Looks like everyone uses moe\_sum in the tidyverse package <https://walker-data.com/census-r/wrangling-census-data-with-tidyverse-tools.html?q=error#calculating-group-wise-margins-of-error>
  + Good datacamp course explaining: <https://campus.datacamp.com/courses/analyzing-us-census-data-in-r/wrangling-us-census-data?ex=8>
  + These functions don’t use the VRTs, which are the more accurate way to calculate the SEs. Especially since Wei 2023 notes that segregation indices tend to
* Basically I need to decide the best way to apply it…
  + Figure out the right sequence in order to calc each part with the variance replicate tables:
    - Low-access:
      * Centroid of block group is in or out of a PD. This doesn’t vary. But the adult pop does vary in summing up the block groups. Put vector of block groups in moe\_sum to add up the adult population and get moe for that.
      * Then ratio of VRTs in order to get proportion of population
    - Low-income:
      * Sums:
      * Proportions:
  + In order to conduct any statistical tests, I need to know the estimate and the associated standard error of each estimate.
  + Instructions for statistical tests with census data : <https://www2.census.gov/programs-surveys/acs/tech_docs/statistical_testing/2019_Instructions_for_Stat_Testing_ACS.pdf>?
    - E.g., significant difference in proportion of individuals in PD vs not with less than a high school education? Need to know estimate and MOE. Calculate the MOE sum when I add up the age groups, calculate the MOE proportion when I divide the n by the N
    - Default is: Standard Error = Margin of Error / 1.645. To change it to a 95% CI, check out Starsinic slides
    - If one of the estimates is a fixed value or comes from a source without sampling error (such as a count from my IIS data later in Aim 1b), use zero for the standard error for that estimate in the Z-score equation.
    - The more estimates I’m adding up, the more different the MOE might be from the real data.
    - Page 3 example: example the proportion of persons 25 and over with a high school diploma or higher.
    - Page 4 example: combining several estimates to create numerator and denominator P = (A + B + C) / (D + E). Before calculating the SE(P), you would first have to find SE(A+B+C) and SE(D+E). You would then use those SEs to calculate SE(P). Can do this all with MOE\_SUM and MOE\_PRODUCT, however it is more accurate to use the VRTs
  + VRTS WEBINAR: <https://www.census.gov/data/academy/webinars/2020/calculating-margins-of-error-acs.html>
  + HOWEVER if VRTS are not available for all the variables we need, we will have to just use the moe\_sum and moe\_product, or at least for the ones without VRTs available.
  + Using the Wei 2023 method, I will basically calculate the end result I need 80 times using the VRTs to get the variance.
    - “we rely upon the VRT to incorporate ACS data uncertainty into segregation measures. Specifically, we compute the segregation measure for each replicate estimate in addition to the published full sample estimate. Then we compute the variance of segregation measures using the same method as the one employed by the ACS during its production.”
    - “Because each set of replicate estimates are estimated using a consistent set of sample weights as mentioned earlier, the correlation or covariance among the same set of replicate estimates is preserved. For example, if the first tract’s white population is positively correlated with the second tract’s white population, this relationship will remain in the replicate estimates even though the individual estimates vary across the 80 sets of replicate estimates.”