# Analyzing US Census Data with tidycensus

Kyle Walker
Texas Christian University
February 24, 2022

# What you'll learn

Part 1: an introduction to **tidycensus**

- Getting started with R and the Census API
- Requesting data from the decennial US Census and American Community Survey
- Customizing Census data outputs
- Analyzing data with **tidyverse** tools
- Visualizing Census data with **ggplot2**

# What you'll learn

Part 2: mapping US Census data in R

- Understanding "simple feature geometry"
- The **tigris** R package and requesting simple feature geometry with **tidycensus**
- Mapping Census data with **tmap**
- Map layouts and interactivity
- Advanced topics: coordinate reference systems and "erasing" water area

# About me

- Associate Professor of Geography at TCU; spatial data science / R consultant
- Book: [*Analyzing US Census Data: Methods, Maps, and Models in R*]
- PhD from Minnesota; undergrad at Oregon (Go Ducks!)
- Twitter: [@kyle_e_walker]

# Workshop setup

- Participants new to R / RStudio: visit https://rstudio.cloud/project/3626443 and sign up for an account (you can authenticate with a Google account). This will set you up with a pre-prepared RStudio environment.

- Experienced users familiar with R / RStudio: clone the workshop repository with `git clone https://github.com/walkerke/uw-workshop.git` and install the required packages in the scripts.

- To get a Census API key, visit https://api.census.gov/data/key_signup.html and follow the instructions.

# Walkthrough: setting up RStudio Cloud

# Part 1: An introduction to tidycensus

# The Census API

- The Census Bureau's main data download interface, data.census.gov, is powered by the Census Application Programming Interface (API)

- The API allows for programmatic access to Census data for developers

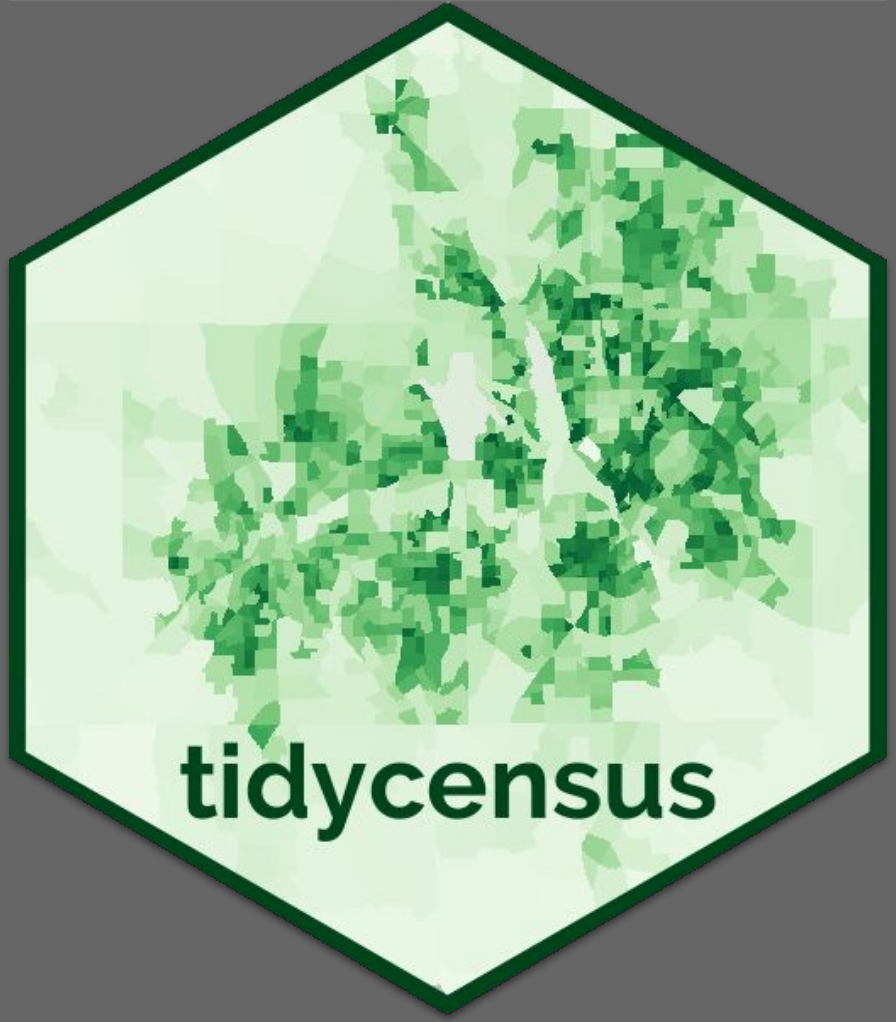
api.census.gov/data/2019/acs/acs5?get=NAME,B01001_001E&for=county:*


```
[["NAME","B01001_001E","state","county"],
["Fayette County, Illinois","21565","17","051"],
["Logan County, Illinois","29003","17","107"],
["Saline County, Illinois","23994","17","165"],
["Lake County, Illinois","701473","17","097"],
["Massac County, Illinois","14219","17","127"],
["Cass County, Illinois","12493","17","017"],
["Huntington County, Indiana","36359","18","069"],
["White County, Indiana","24149","18","181"],
["Jay County, Indiana","20840","18","075"],
["Shelby County, Indiana","44438","18","145"],
["Sullivan County, Indiana","20730","18","153"],
["Tippecanoe County, Indiana","191553","18","157"],
["Hamilton County, Indiana","323117","18","057"],
["Bartholomew County, Indiana","82481","18","005"],
["Fulton County, Indiana","20096","18","049"],
["Noble County, Indiana","47506","18","113"],
["Clark County, Indiana","116507","18","019"],
["Hendricks County, Indiana","163799","18","063"],
["Grant County, Indiana","66452","18","053"],
["Jackson County, Indiana","44025","18","071"],
["Owen County, Indiana","20835","18","119"],
["Whitley County, Indiana","33730","18","183"],
["Clinton County, Indiana","32273","18","023"],
["Union County, Indiana","7113","18","161"],
["Dearborn County, Indiana","49479","18","029"],
["Lawrence County, Indiana","45548","18","093"],
["Perry County, Indiana","19102","18","123"],
["Posey County, Indiana","25560","18","129"],
["Carroll County, Indiana","20074","18","015"],
["Fountain County, Indiana","16430","18","045"],
["Starke County, Indiana","22952","18","149"],
```

# The tidycensus R package

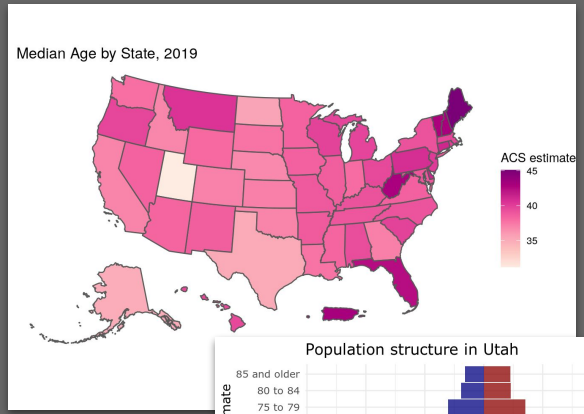
- First released in 2017, **tidycensus** is an R package that helps users acquire pre-formatted Census data

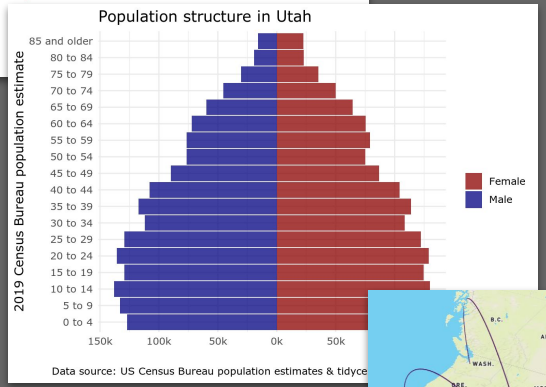
- Over 234,000 downloads from the RStudio CRAN mirror since its debut

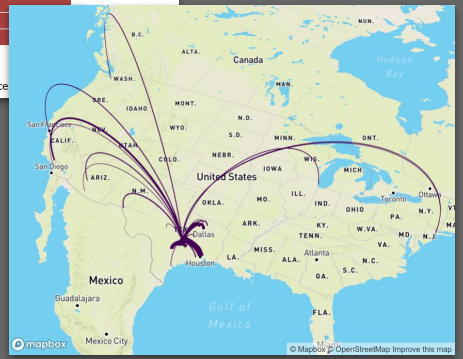


# The tidycensus R package

- Original motivation: streamline the process of getting decennial Census / ACS data with geometry (GIS data) pre-joined
- Core functions: `get_decennial()` and `get_acs()`
- The project has since evolved to accommodate ACS microdata, Census population estimates, and migration flows data as well

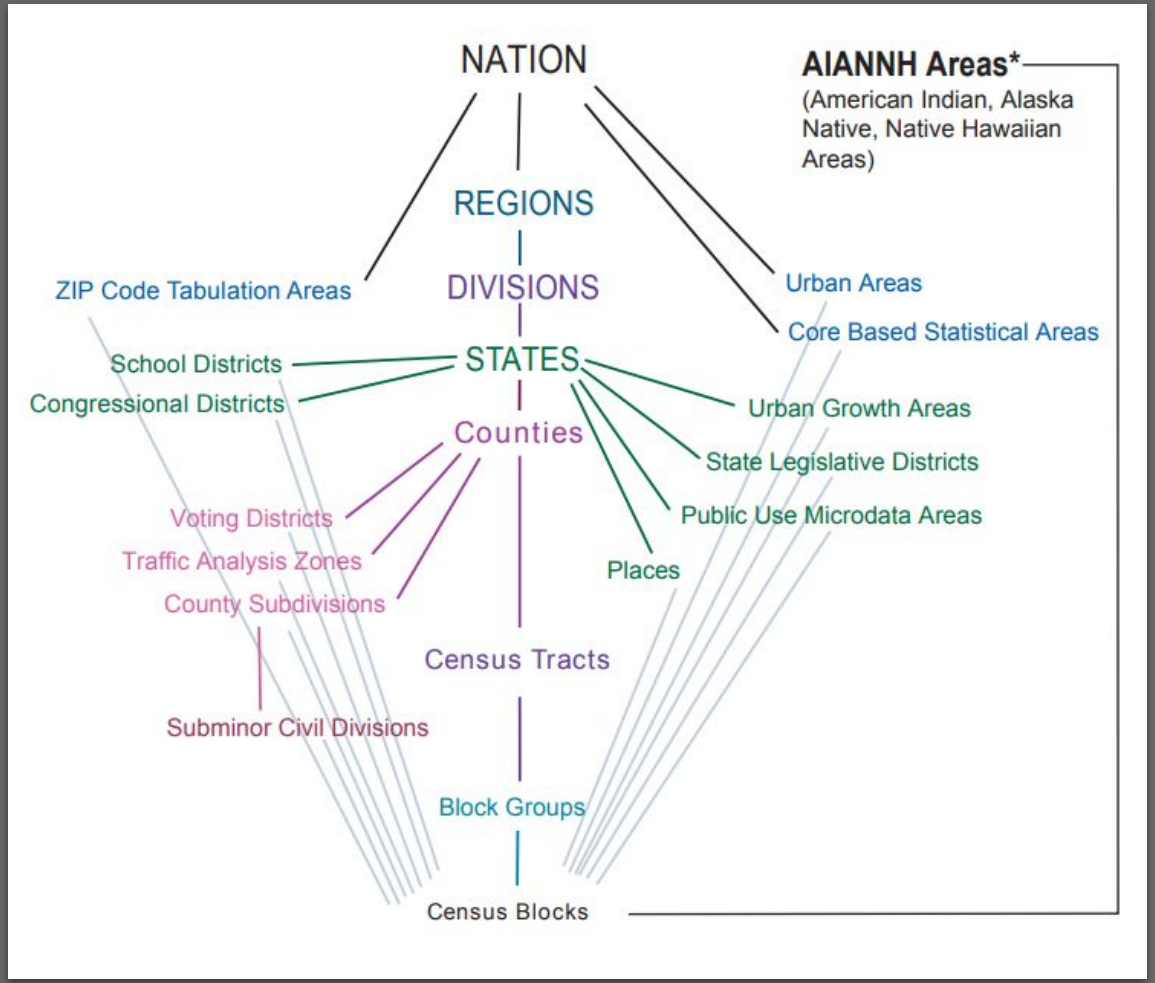


# Demo: getting started with tidycensus

# Census geography



# Census variables and datasets

- While hundreds of datasets are available from the Census API (see the **censusapi** R package!), **tidycensus** focuses on a select few, centered around the decennial Census and ACS

- To look up variables for a given dataset, use the `load_variables()` function and then browse the result in RStudio with `View()`

| name | label | concept |
|---|---|---|
| B01001_001 | Estimate!!Total: | SEX BY AGE |
| B01001_002 | Estimate!!Total:!!Male: | SEX BY AGE |
| B01001_003 | Estimate!!Total:!!Male:!!Under 5 years | SEX BY AGE |
| B01001_004 | Estimate!!Total:!!Male:!!5 to 9 years | SEX BY AGE |
| B01001_005 | Estimate!!Total:!!Male:!!10 to 14 years | SEX BY AGE |
| B01001_006 | Estimate!!Total:!!Male:!!15 to 17 years | SEX BY AGE |
| B01001_007 | Estimate!!Total:!!Male:!!18 and 19 years | SEX BY AGE |
| B01001_008 | Estimate!!Total:!!Male:!!20 years | SEX BY AGE |
| B01001_009 | Estimate!!Total:!!Male:!!21 years | SEX BY AGE |
| B01001_010 | Estimate!!Total:!!Male:!!22 to 24 years | SEX BY AGE |
| B01001_011 | Estimate!!Total:!!Male:!!25 to 29 years | SEX BY AGE |
| B01001_012 | Estimate!!Total:!!Male:!!30 to 34 years | SEX BY AGE |
| B01001_013 | Estimate!!Total:!!Male:!!35 to 39 years | SEX BY AGE |
| B01001_014 | Estimate!!Total:!!Male:!!40 to 44 years | SEX BY AGE |
| B01001_015 | Estimate!!Total:!!Male:!!45 to 49 years | SEX BY AGE |
| B01001_016 | Estimate!!Total:!!Male:!!50 to 54 years | SEX BY AGE |
| B01001_017 | Estimate!!Total:!!Male:!!55 to 59 years | SEX BY AGE |
| B01001_018 | Estimate!!Total:!!Male:!!60 and 61 years | SEX BY AGE |
| B01001_019 | Estimate!!Total:!!Male:!!62 to 64 years | SEX BY AGE |
| B01001_020 | Estimate!!Total:!!Male:!!65 and 66 years | SEX BY AGE |
| B01001_021 | Estimate!!Total:!!Male:!!67 to 69 years | SEX BY AGE |
| B01001_022 | Estimate!!Total:!!Male:!!70 to 74 years | SEX BY AGE |
| B01001_023 | Estimate!!Total:!!Male:!!75 to 79 years | SEX BY AGE |
| B01001_024 | Estimate!!Total:!!Male:!!80 to 84 years | SEX BY AGE |
| B01001_025 | Estimate!!Total:!!Male:!!85 years and over | SEX BY AGE |

# tidycensus data structure

**tidycensus** users can request data in two output formats:

- `output = "tidy"` (the default), which returns data in *long form*;
- `output = "wide"`, with Census variables spread across the columns

| GEOID | NAME | variable | estimate | moe |
|---|---|---|---|---|
| 53001 | Adams County, Washington | B19001_001 | 5973 | 159 |
| 53001 | Adams County, Washington | B19001_002 | 490 | 147 |
| 53001 | Adams County, Washington | B19001_003 | 233 | 96 |
| 53001 | Adams County, Washington | B19001_004 | 294 | 107 |
| 53001 | Adams County, Washington | B19001_005 | 463 | 150 |
| 53001 | Adams County, Washington | B19001_006 | 331 | 140 |
| 53001 | Adams County, Washington | B19001_007 | 203 | 68 |
| 53001 | Adams County, Washington | B19001_008 | 317 | 120 |
| 53001 | Adams County, Washington | B19001_009 | 479 | 164 |
| 53001 | Adams County, Washington | B19001_010 | 227 | 102 |
| 53001 | Adams County, Washington | B19001_011 | 439 | 118 |
| 53001 | Adams County, Washington | B19001_012 | 706 | 163 |
| 53001 | Adams Cou | | | |
| 53001 | Adams Cou | | | |
| 53001 | Adams Cou | | | |

| GEOID | NAME | B19001_001E | B19001_001M | B19001_002E | B19001_002M |
|---|---|---|---|---|---|
| 53001 | Adams County, Washington | 5973 | 159 | 490 | 147 |
| 53003 | Asotin County, Washington | 9101 | 231 | 382 | 120 |
| 53005 | Benton County, Washington | 72121 | 607 | 2999 | 442 |
| 53007 | Chelan County, Washington | 28384 | 668 | 1536 | 350 |
| 53009 | Clallam County, Washington | 32958 | 517 | 1948 | 275 |
| 53011 | Clark County, Washington | 174661 | 837 | 6070 | 545 |
| 53013 | Columbia County, Washington | 1795 | 99 | 112 | 41 |
| 53015 | Cowlitz County, Washington | 41952 | 474 | 2729 | 383 |
| 53017 | Douglas County, Washington | 15263 | 249 | 560 | 133 |
| 53019 | Ferry County, Washington | 3060 | 164 | 295 | 87 |
| 53021 | Franklin County, Washington | 26723 | 310 | 1112 | 301 |
| 53023 | Garfield County, Washington | 984 | 72 | 63 | 45 |
| 53025 | Grant County, Washington | 30818 | 612 | 1559 | 308 |
| 53027 | Grays Harbor County, Washington | 28722 | 648 | 2086 | 401 |
| 53029 | Island County, Washington | 34768 | 566 | 1679 | 278 |

# Demo: geography, variables, and tables in tidycensus

# tidycensus and the tidyverse

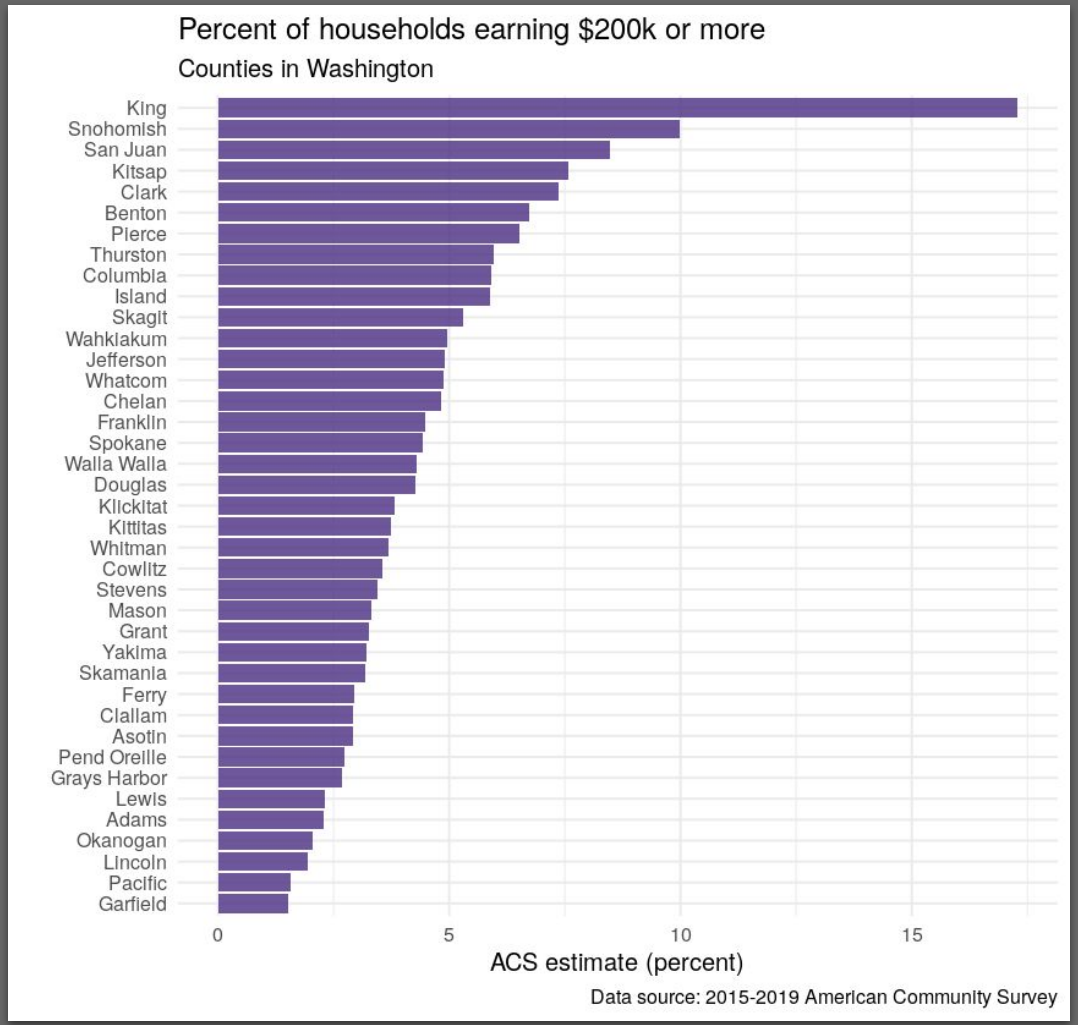
- The **tidyverse**: a popular suite of packages for exploratory data analysis / data wrangling maintained by RStudio

- **tidycensus** was originally developed to help analysts work with Census data in a **tidyverse**-friendly way

# Visualizing US Census data

- Census data can be used to create a wide range of compelling visualizations - see [Chapter 4 of *Analyzing US Census Data*](#)!

- Let's walk through how to use **tidyverse** tools to create the graphic to the right



# Exercises

1. Look up variables from the 2020 decennial Census with `load_variables(2020, "pl")`.

2. Get a dataset with information on total population and the Hispanic population from the 2020 decennial Census for counties in Washington using appropriate Census variable IDs and `get_decennial()`.

3. Bonus: try to adapt the code above to find out which county in Washington has the largest Hispanic share of its population in 2020.  We'll review after the break!

# Part 2: mapping Census data

# Exercises #1 and #2

```
library(tidycensus)
library(tidyverse)

vars2020 <- load_variables(2020, "pl")

wa_2020 <- get_decennial(
  geography = "county",
  state = "WA",
  variables = c("P2_001N", "P2_002N"),
  year = 2020,
  output = "wide"
)
```

# Exercise #3

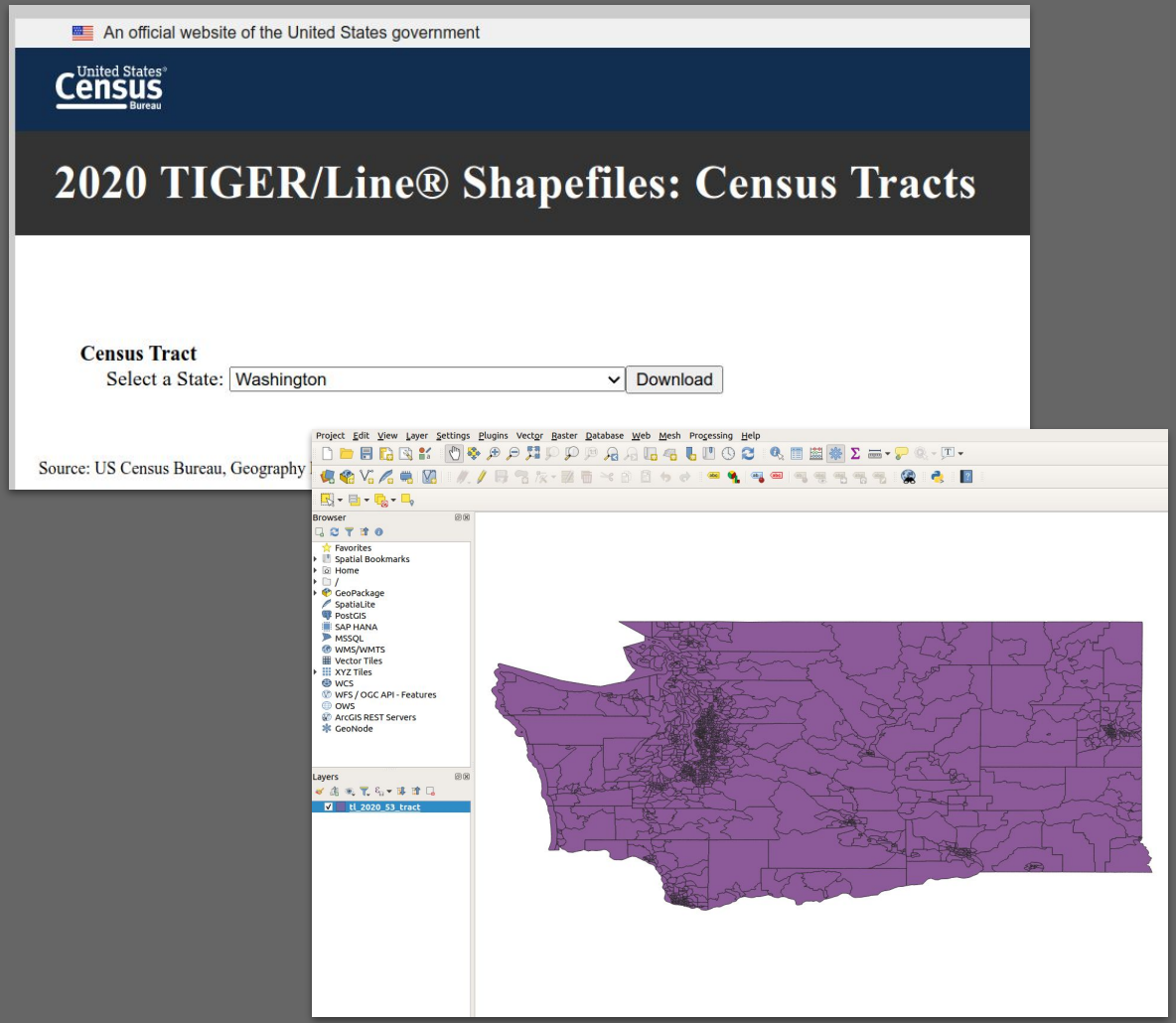
```
library(tidycensus)
library(tidyverse)

# Use mutate() to create a new column and arrange() to view
# the top values
wa_pct_hispanic <- wa_2020 %>%
  mutate(percent_hispanic = 100 * (P2_002N / P2_001N)) %>%
  arrange(desc(percent_hispanic))
```

# Typical GIS workflows

- The US Census Bureau releases extracts from its TIGER/Line database as *shapefiles*, which can be downloaded from the Census website

- These shapefiles are typically opened in dedicated GIS software like ArcGIS or QGIS



# The tigris R package

- I first developed **tigris** in 2015 to automate loading of Census shapefiles into R and avoid navigating menus / remembering FIPS codes

- Since then: downloaded over 400,000 times from the RStudio CRAN mirror

- Available datasets include Census enumeration areas (counties, tracts) as well as geographic features like roads and water areas

# Census geometries in R

- **tigris** returns Census geographic data as "simple feature geometry", implemented in the **sf** R package

- Census shapes typically come from the core TIGER/Line shapefiles (the default) or the cartographic boundary shapefiles with the argument `cb = TRUE`

- **tidycensus** users can get Census data pre-joined with geometry automatically!

| GEOID | NAME | total_pop | hispanic | geometry |
|---|---|---|---|---|
| 53055 | San Juan County, Washington | 17788 | 1298 | *list(list(c(-122.766567, -122.765633, -122.764552, -...* |
| 53031 | Jefferson County, Washington | 32977 | 1305 | *list(list(c(-122.94347, -122.929488, -122.916706, -1...* |
| 53015 | Cowlitz County, Washington | 110730 | 10802 | *list(list(c(-123.218309, -123.21795, -123.176508, -1...* |
| 53053 | Pierce County, Washington | 921130 | 111811 | *list(list(c(-122.640638, -122.638649, -122.635884, -...* |
| 53061 | Snohomish County, Washington | 827957 | 95644 | *list(list(c(-122.33164, -122.328343, -122.322362, -1...* |
| 53019 | Ferry County, Washington | 7178 | 210 | *list(list(c(-118.869633, -118.837006, -118.836999, -...* |
| 53021 | Franklin County, Washington | 96749 | 52445 | *list(list(c(-119.456993, -119.45319, -119.432409, -1...* |
| 53073 | Whatcom County, Washington | 226847 | 22825 | *list(list(c(-122.593346, -122.587158, -122.586678, -...* |
| 53067 | Thurston County, Washington | 294793 | 29024 | *list(list(c(-123.200888, -123.201013, -123.200041, -...* |
| 53071 | Walla Walla County, Washington | 62584 | 14206 | *list(list(c(-119.039918, -119.037218, -119.031961, -...* |
| 53045 | Mason County, Washington | 65726 | 7595 | *list(list(c(-123.505916, -123.505917, -123.505923, -...* |
| 53029 | Island County, Washington | 86857 | 7118 | *list(list(c(-122.538916, -122.538234, -122.534431, -...* |
| 53057 | Skagit County, Washington | 129523 | 23792 | *list(list(c(-122.537676576494, -122.535835, -122.5...* |
| 53011 | Clark County, Washington | 503311 | 58790 | *list(list(c(-122.795963, -122.785696, -122.785515, -...* |
| 53005 | Benton County, Washington | 206873 | 49339 | *list(list(c(-119.875084, -119.874337, -119.874042, -...* |

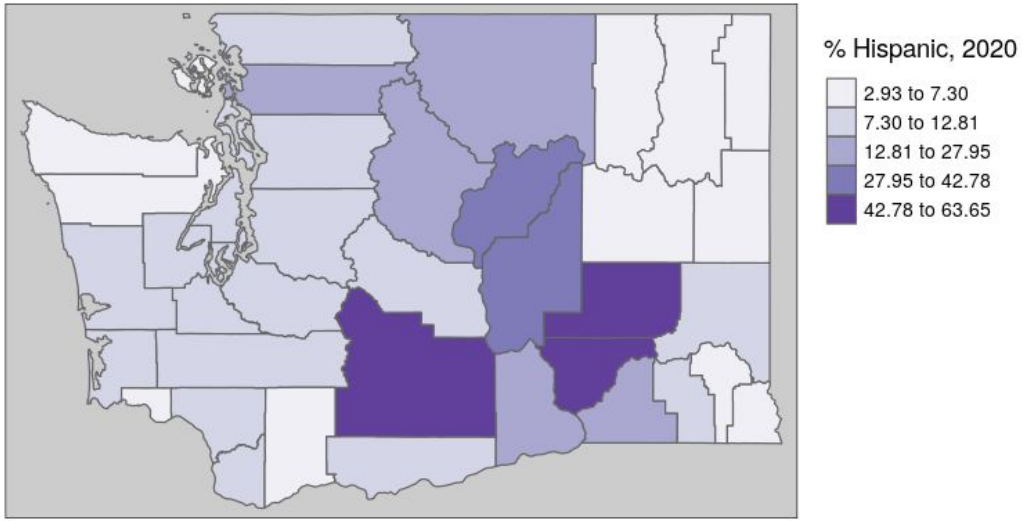
# Demo: Census geometries in R

# Mapping Census data in R

- The **ggplot2** and **tmap** R packages are the most popular packages for mapping spatial data

- Analysts familiar with mapping in a desktop GIS will likely find **tmap** intuitive

- Maps from data are initialized with `tm_shape()`, then cartographic design and layout elements are layered on

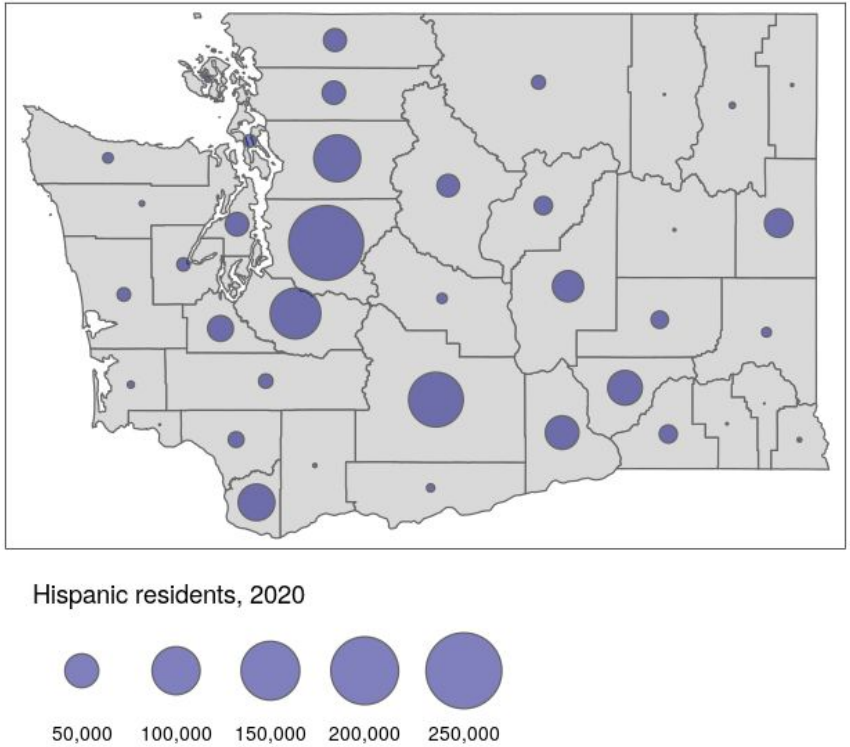


# Customizing map outputs with tmap

- **tmap** includes a wide range of options for customization of maps, including ColorBrewer color palettes and customizable breaks (e.g. Jenks natural breaks) familiar to GIS users

- Alternative map types including graduate symbol and dot-density maps are also available with **tmap**

- Convert to an interactive map with `tmap_mode("view")`; switch back with `tmap_mode("plot")`

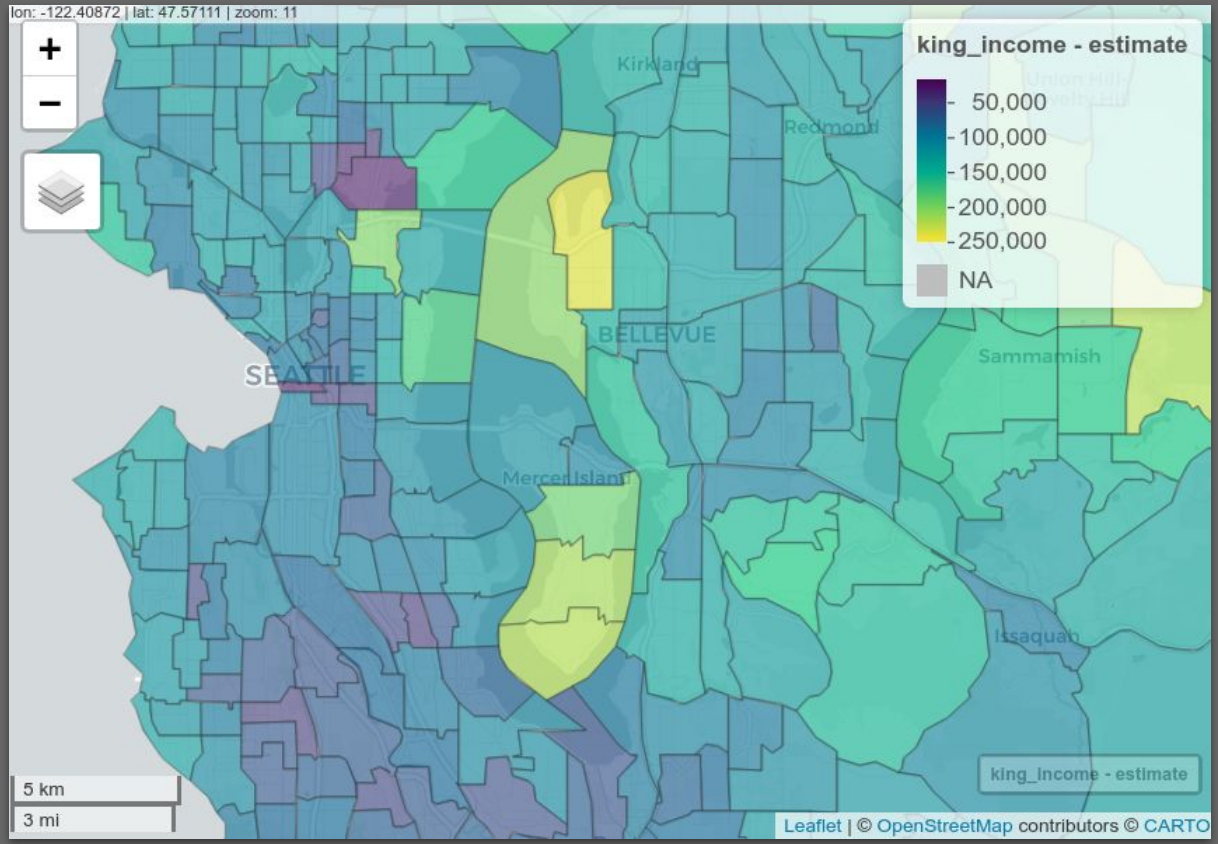


# Demo: making maps in R with tmap

# Advanced workflow: modifying geometries

- For water-rich areas like the Seattle region, shapefiles available from the Census Bureau (even the cartographic boundary files) may be insufficiently detailed

- For example: where is Mercer Island?



# Choosing a coordinate reference system

- While the **sf** R package has functionality for working with spherical geometries, it is often faster to work in a projected coordinate reference system

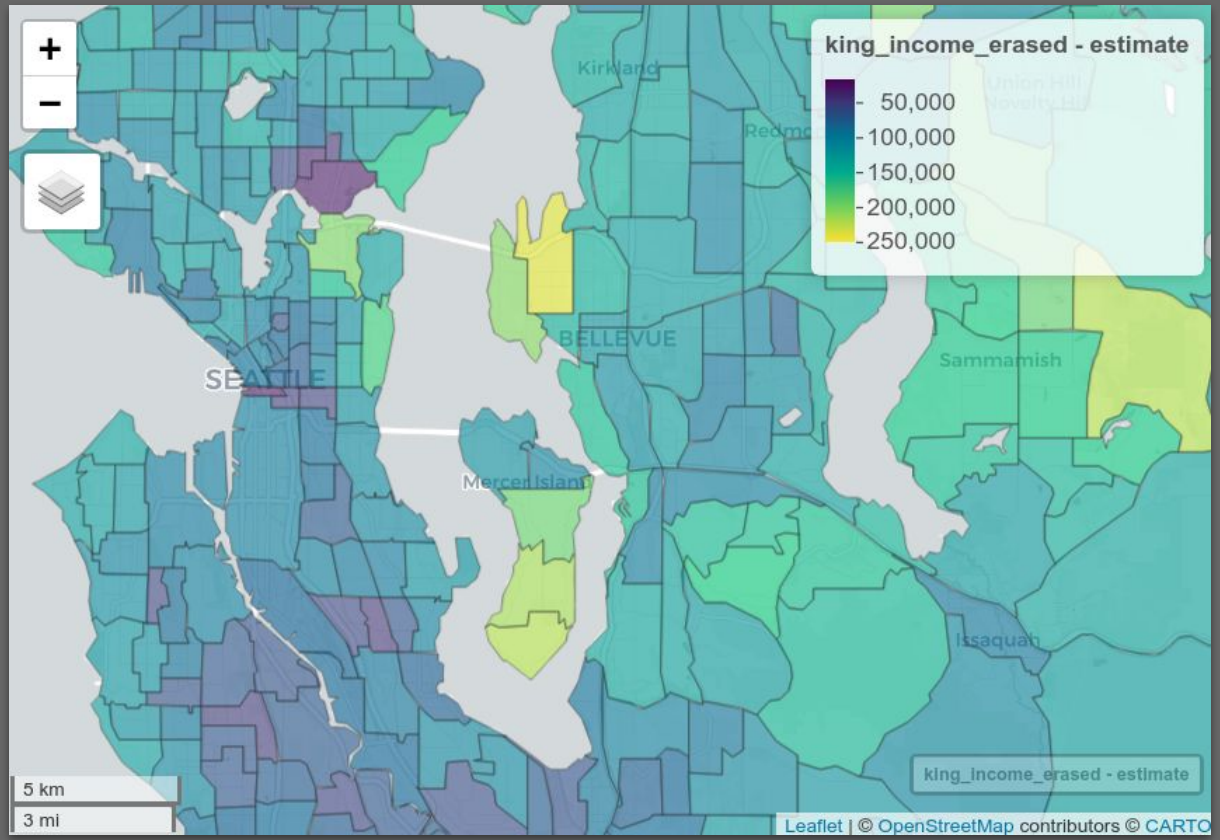
- The **crsuggest** R package helps us choose an appropriate CRS, which can be used in a CRS transformation with the `st_transform()` function

| crs_code | crs_name | crs_type | crs_gcs | crs_units |
|---|---|---|---|---|
| 6597 | NAD83(2011) / Washington North (ftUS) | projected | 6318 | us-ft |
| 6596 | NAD83(2011) / Washington North | projected | 6318 | m |
| 3690 | NAD83(NSRS2007) / Washington North (ftUS) | projected | 4759 | us-ft |
| 3689 | NAD83(NSRS2007) / Washington North | projected | 4759 | m |
| 32148 | NAD83 / Washington North | projected | 4269 | m |
| 32048 | NAD27 / Washington North | projected | 4267 | us-ft |
| 2926 | NAD83(HARN) / Washington North (ftUS) | projected | 4152 | us-ft |
| 2855 | NAD83(HARN) / Washington North | projected | 4152 | m |
| 2285 | NAD83 / Washington North (ftUS) | projected | 4269 | us-ft |
| 6599 | NAD83(2011) / Washington South (ftUS) | projected | 6318 | us-ft |

# Brand-new feature: `erase_water()`

- I just added the erase_water() function to **tigris**, which automates the process of removing water area from geometries in areas like the Seattle region

- Removal of large bodies of water from Census polygons (like Lake Washington) can dramatically improve cartographic displays



# Demo: advanced geometry workflows with Census data in R

# Exercises

1. Use your knowledge gained in both parts of this workshop to get spatial Census data for a different location and different Census variable with **tidycensus**.

2. Make a static map of that data with **tmap**.

3. Make an interactive map of that data with **tmap** as well!

# Thank you!