NYU DataServices

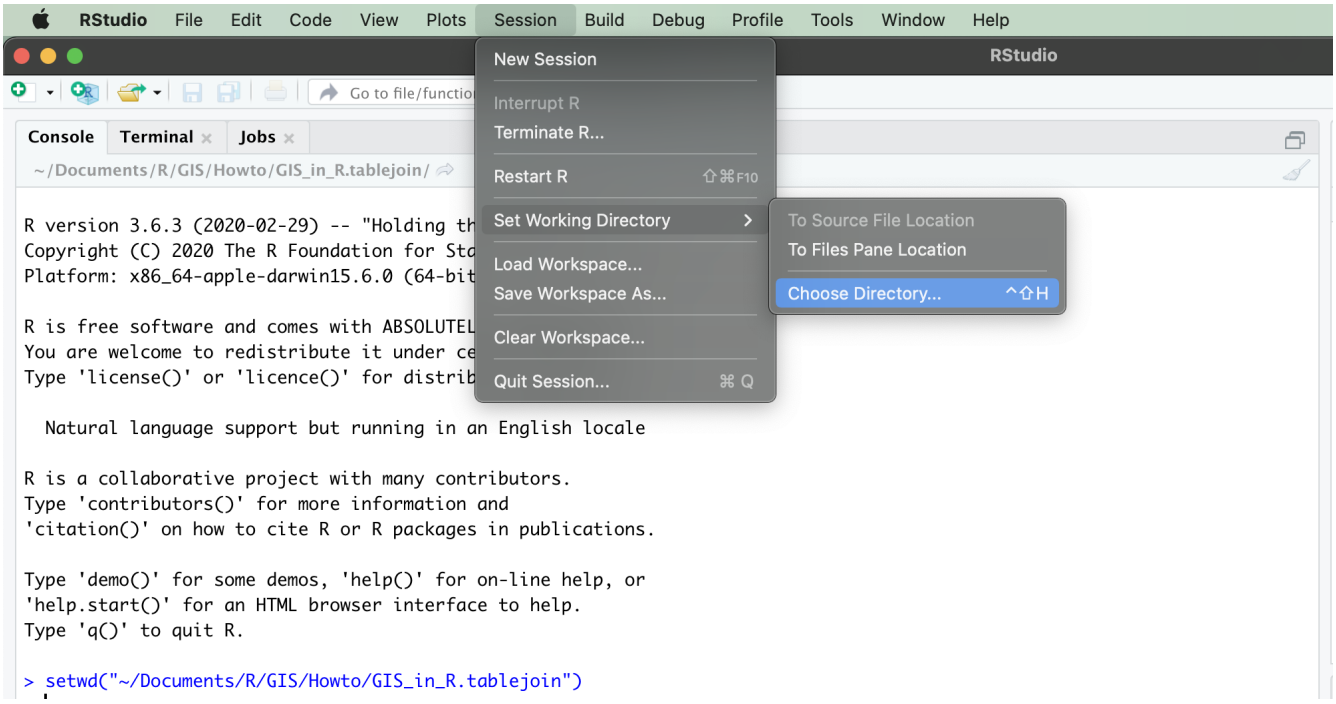
How to join tabular(CSV) data to a shapefile

GIS in R: Table Join

Joining Tabular data to a shapefile allows you to add fields from your data table to a shapefile's attribute table based on matching values found in the key columns, so that you may generate a choropleth map of your variable of interest. You can use 'st_read' function that is part of the sf package to read the shapefile into R, then use join functions in dplyr package to merge the tabular data with shapefile based on common columns, and use tmap to show and format the choropleth map of the interested variable.

For example, as a number of research show that people of color are far more likely to be uninsured in America, you would like to study in which states do African Americans have high rates of uninsurance. However, you only have a U.S. states shapefile without the attribute of health insurance and a tabular file inlcuding the insurance data at state-level downloaded from U.S. Census. Also, you would like to conduct the analysis in R.

1. Set your working directory to the folder where your project locates using `setwd()`. Alternatively, you can also set the working directory using the dropdown menu **Session** in RStudio. Select the **Choose Directory** and then click the project folder. After you've set the working directory, you may verify it by calling the `getwd()` function.



2. Load the following packages in R. If you do not already have the packages installed, be sure to install them using command `install.packages()` before loading them.

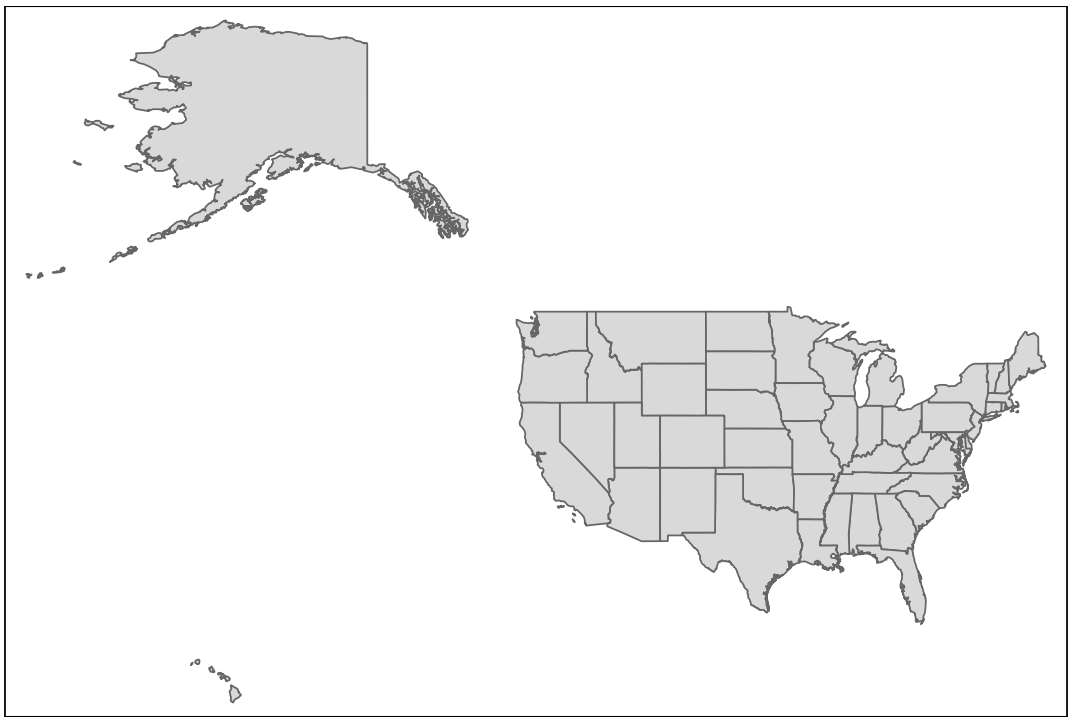
```
library(dplyr)
library(sf)
library(tmap)
```

3. Use the `st_read()` function to read the shapefile into R. Optionally, you can use the mapping and spatial data visualization package called tmap to generate a simple outline of the shapefile. The `tm_shape()` function specifies the file you wish to map, while the `tm_polygons()` function tells tmap to draw the polygons that comprise this shapefile. Note it is also possible to use different ways to call forth a visual representation of the shapefile (e.g. by using ggplot).

```
US_shapefile <- st_read("nyu_2451_34094.shp",
                        stringsAsFactors = FALSE)
```

```
## Reading layer `nyu_2451_34094' from data source `/Users/junhui/Documents/R/GIS/Howto/GIS_in_R.tablejoin
## Simple feature collection with 51 features and 50 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -178.2176 ymin: 18.92179 xmax: -66.96927 ymax: 71.40624
## CRS:            4326
```

```
tm_shape(US_shapefile) + tm_polygons()
```



4. Read the tabular data and examine the structure of the dataset. Make sure the tabular dataset has at least one column same as in the shapefile that you can use as key column to join the two. You may also want to use `str()` to check if the key columns in the two datasets are of same type, and if the type of the variable of interst is desired. For this example, both US_shapefile and uninsurance_rate have column "STATE_NAME" of type "character", and the column "Black.or.African.American.alone" is of type "numeric".

```
uninsurance_rate <- read.csv("UninsuranceRate.csv",
                             stringsAsFactors = FALSE)
# str(uninsurance_rate)
```

5. Once you've had shapefile and tabular data read into R and ready in format, you can use join functions to merge the two datasets based on the field that is common to both.

```
US_uninsured_shapefile <- left_join(US_shapefile, uninsurance_rate, by = "STATE_NAME")
```

6. You can now use tmap to generate a choropleth map of your variable of interest. For this example, based on the "US_uninsured_shapefile" we generated in Step 2, further tell tmap that you want to map the values contained in the column "Black.or.African.American.alone", that we want the data to be partitioned into 6 bins (n=6), that we want to set break points in the data distribution using the Jenks Natural Breaks Classification (style="jenks"), and that we want a Yellow/Red color palette (palette="YlOrRd"). In the choropleth map below, you may see that the state of highest percentage of Black and African American without health insurance in 2019 is North Dakota .

```
map <- tm_shape(US_uninsured_shapefile) +
  tm_polygons(col="Black.or.African.American.alone", n=6,
              title="Uninsured % of \nBlack or African American",
              style="jenks", palette="YlOrRd")
map
```