In-class worksheet 29

May 7, 2019

In this worksheet, we will use the libraries tidyverse and sf:

```
library(tidyverse)
theme_set(theme_bw(base_size=12)) # set default ggplot2 theme
library(sf) # needed for simple feature manipulation
```

1. Manipulating and plotting geospatial data

We will work with two data frames, `US_income` and `US_counties_income`, which contain the median income and population number of US states or US counties, respectively.

```
# load all data
load(url("https://wilkelab.org/classes/SDS348/data_sets/US_income.RData"))

# income and population data by state
head(US_income)
```

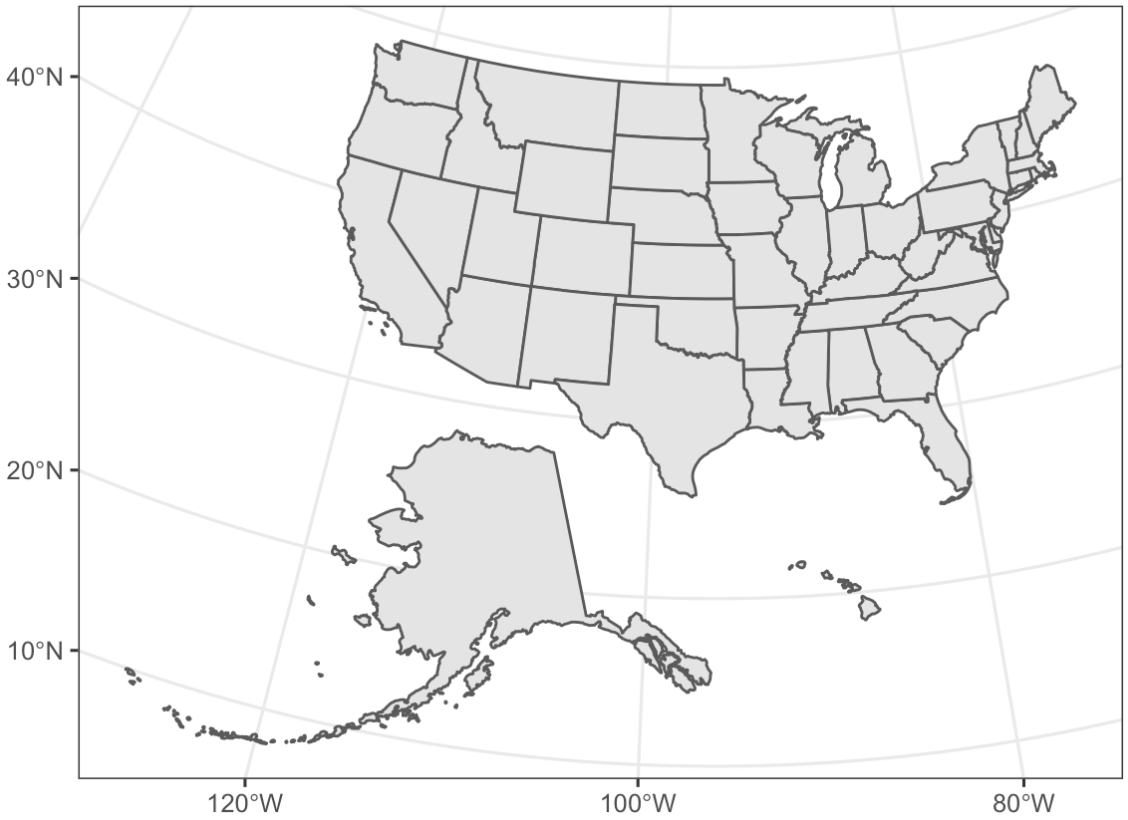
```
## Simple feature collection with 6 features and 7 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -2356114 ymin: -778242.8 xmax: 1986024 ymax: 845925.2
## epsg (SRID):    NA
## proj4string:    +proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=37.5 +lon_0=-96 +x_
0=0 +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
##   GEOID        name median_income median_income_moe population
## 1    01     Alabama         43623               281    4830620
## 2    04     Arizona         50255               211    6641928
## 3    05    Arkansas         41371               247    2958208
## 4    06  California         61818               156   38421464
## 5    08    Colorado         60629               252    5278906
## 6    09 Connecticut         70331               409    3593222
##                 area              popdens                       geometry
## 1 133958437749 [m^2] 3.606059e-05 [1/m^2] MULTIPOLYGON (((1032679 -63...
## 2 295232708152 [m^2] 2.249726e-05 [1/m^2] MULTIPOLYGON (((-1216674 -4...
## 3 137792577218 [m^2] 2.146856e-05 [1/m^2] MULTIPOLYGON (((462619.4 -3...
## 4 410516610493 [m^2] 9.359296e-05 [1/m^2] MULTIPOLYGON (((-2077630 -2...
## 5 269580118211 [m^2] 1.958196e-05 [1/m^2] MULTIPOLYGON (((-527710.6 3...
## 6  12961831628 [m^2] 2.772156e-04 [1/m^2] MULTIPOLYGON (((1841099 622...
```

```
# income and population data by county
head(US_counties_income)
```

```
## Simple feature collection with 6 features and 15 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -2284310 ymin: -146995 xmax: 2024652 ymax: 1066541
## epsg (SRID):    NA
## proj4string:    +proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=37.5 +lon_0=-96 +x_
0=0 +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
##   STATEFP COUNTYFP COUNTYNS       AFFGEOID GEOID          NAME LSAD
## 1      06      075 00277302 0500000US06075 06075 San Francisco   06
## 2      25      025 00606939 0500000US25025 25025       Suffolk   06
## 3      31      007 00835826 0500000US31007 31007        Banner   06
## 4      37      181 01008591 0500000US37181 37181         Vance   06
## 5      48      421 01383996 0500000US48421 48421       Sherman   06
## 6      50      011 01461762 0500000US50011 50011      Franklin   06
##        ALAND    AWATER                             name median_income
## 1  121485107 479107241 San Francisco County, California         81294
## 2  150855462 160479920    Suffolk County, Massachusetts         55044
## 3 1932676697    397069          Banner County, Nebraska         48897
## 4  653705784  42187365     Vance County, North Carolina         33316
## 5 2390651189    428754            Sherman County, Texas         51987
## 6 1641633748 150930318         Franklin County, Vermont         58199
##   median_income_moe population             area             popdens
## 1              1099     840763  113979848 [m^2] 7.376418e-03 [1/m^2]
## 2               992     758919  180163309 [m^2] 4.212395e-03 [1/m^2]
## 3              4107        820 1926477562 [m^2] 4.256473e-07 [1/m^2]
## 4              1974      44829  673640216 [m^2] 6.654739e-05 [1/m^2]
## 5              4386       3066 2387929738 [m^2] 1.283957e-06 [1/m^2]
## 6              2034      48418 1798183349 [m^2] 2.692606e-05 [1/m^2]
##                         geometry
## 1 MULTIPOLYGON (((-2283315 35...
## 2 MULTIPOLYGON (((2009657 799...
## 3 MULTIPOLYGON (((-664543.3 4...
## 4 MULTIPOLYGON (((1544259 321...
## 5 MULTIPOLYGON (((-546533.1 -...
## 6 MULTIPOLYGON (((1780210 102...
```

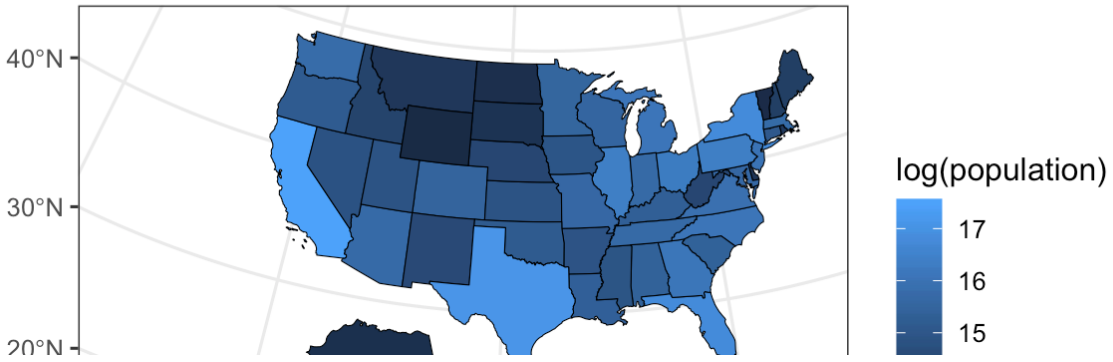
Both data frames contain the boundaries of the lower 48 states as well as Alaska and Hawaii. For easier visualization, Alaska and Hawaii have been moved to lie underneath the lower 48 states. We can plot the geographic boundaries with `geom_sf()` ("sf" stands for "simple features"). Note that for a basic plot, we don't need to specify an aesthetic mapping, because geometry columns are automatically found and mapped by `geom_sf()`.

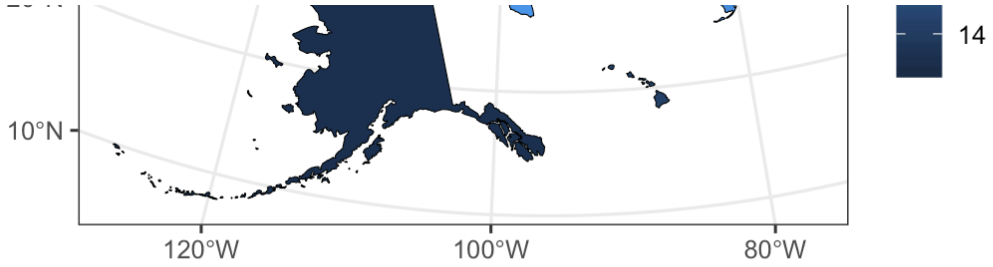
```
ggplot(US_income) +
  geom_sf()
```



We can map any of the other data values onto the map using standard ggplot2 techniques. For example, we can color states by the logarithm of the population number.

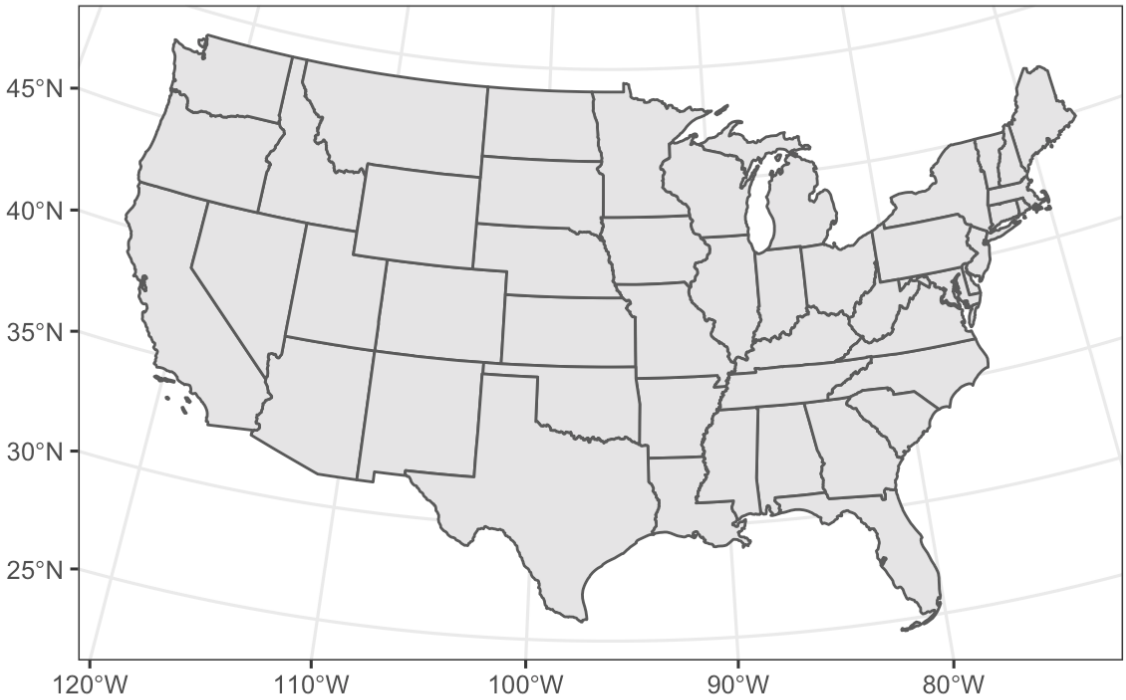
```
ggplot(US_income, aes(fill = log(population))) +
  geom_sf(color = "black", size = 0.2) # draw state boundaries with thin black
lines
```





We can remove states we are not interested in by filtering, just like we normally do when working with the tidyverse.

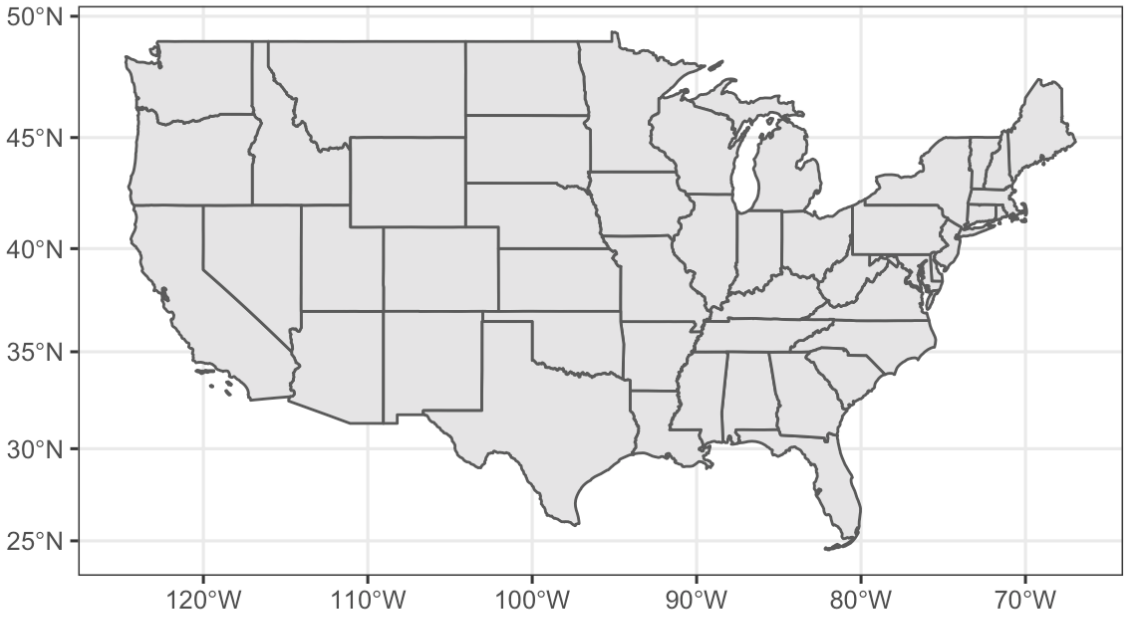
```
# remove Alaska and Hawaii
lower48 <- US_income %>%
  filter(!GEOID %in% c("02", "15"))

# plot
ggplot(lower48) + geom_sf()
```



We can change the coordinate system (i.e., reproject the geometric shapes) by adding `coord_sf()` with a coordinate reference system (crs). Many coordinate reference systems are specified by EPSG

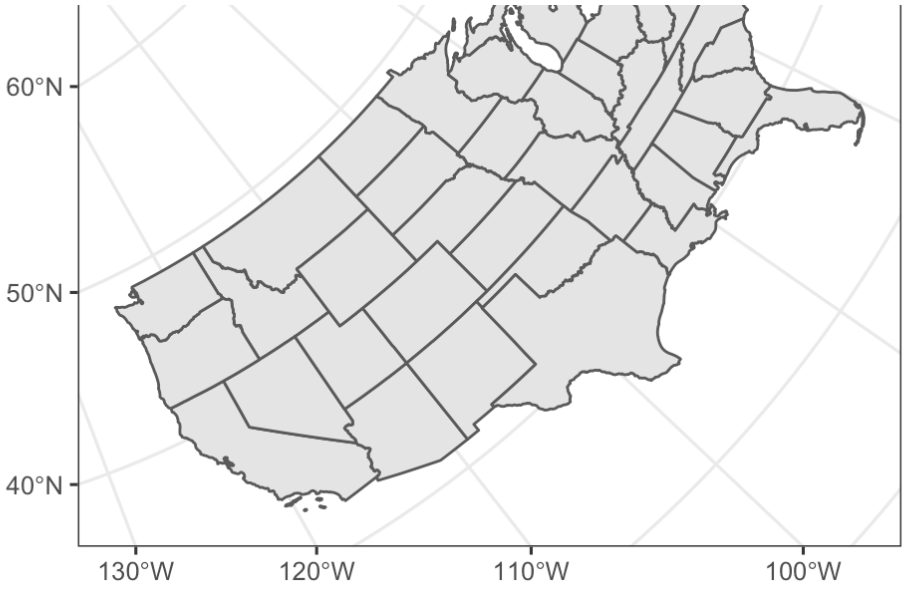
(European Petroleum Survey Group) codes, which can be looked up at https://epsg.io/ (https://epsg.io/) or https://spatialreference.org (https://spatialreference.org). For example, we can use EPSG 3395, which is an outdated Mercator projection.

```
ggplot(lower48) +
  geom_sf() +
  coord_sf(crs = 3395) # World Mercator, not recommended in practice, https://spatialreference.org/ref/epsg/3395/
```



Or, we could use EPSG 3338, which is a projection that is normally used for Alaska.

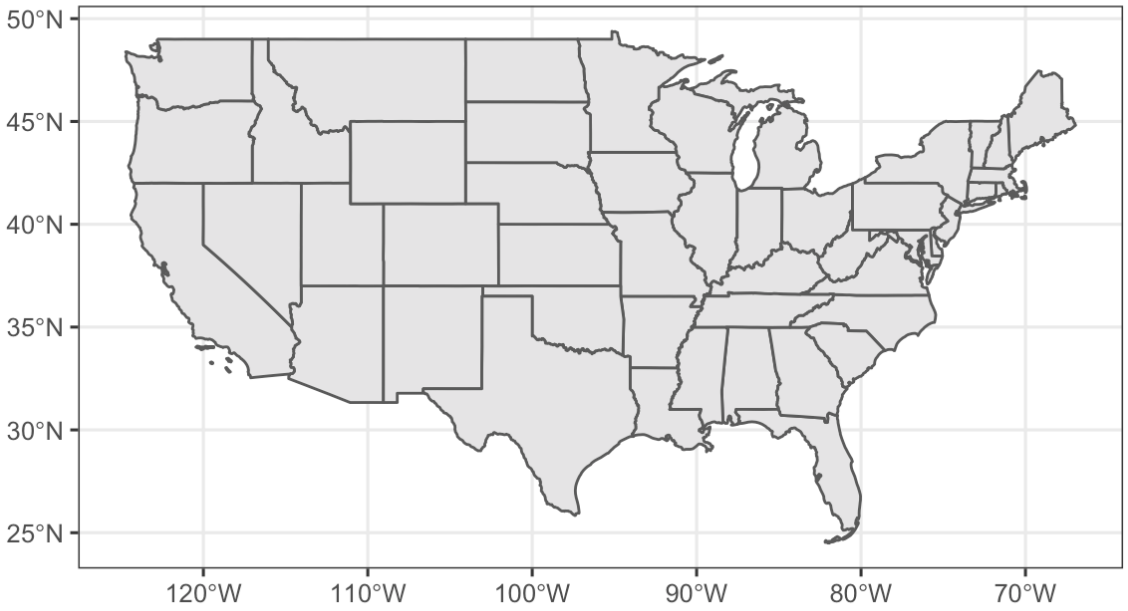
```
ggplot(lower48) +
  geom_sf() +
  coord_sf(crs = 3338) # Normally used for Alaska, https://spatialreference.org/ref/epsg/nad83-alaska-albers/
```



2. Problems

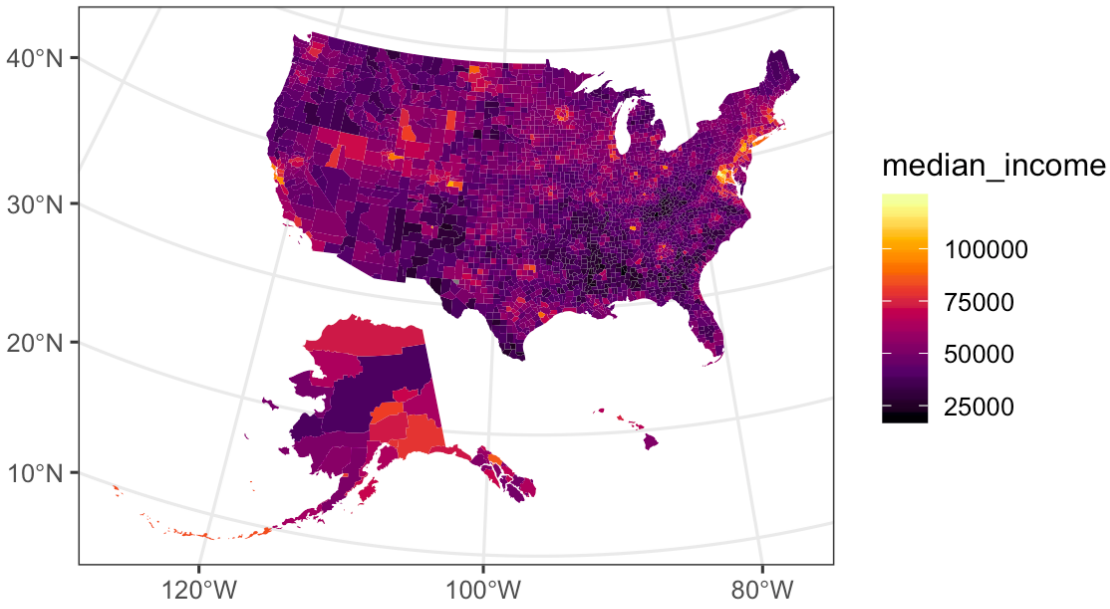
Plot the lower 48 states in a coordinate system that represents longitude along the x axis and latitude along the y axis. Hint: This is called the longitude/latitude projection, and it has an EPSG code of 4326.

```
ggplot(lower48) +
  geom_sf() +
  coord_sf(crs = 4326) # Cartesian longitude and latitude
```



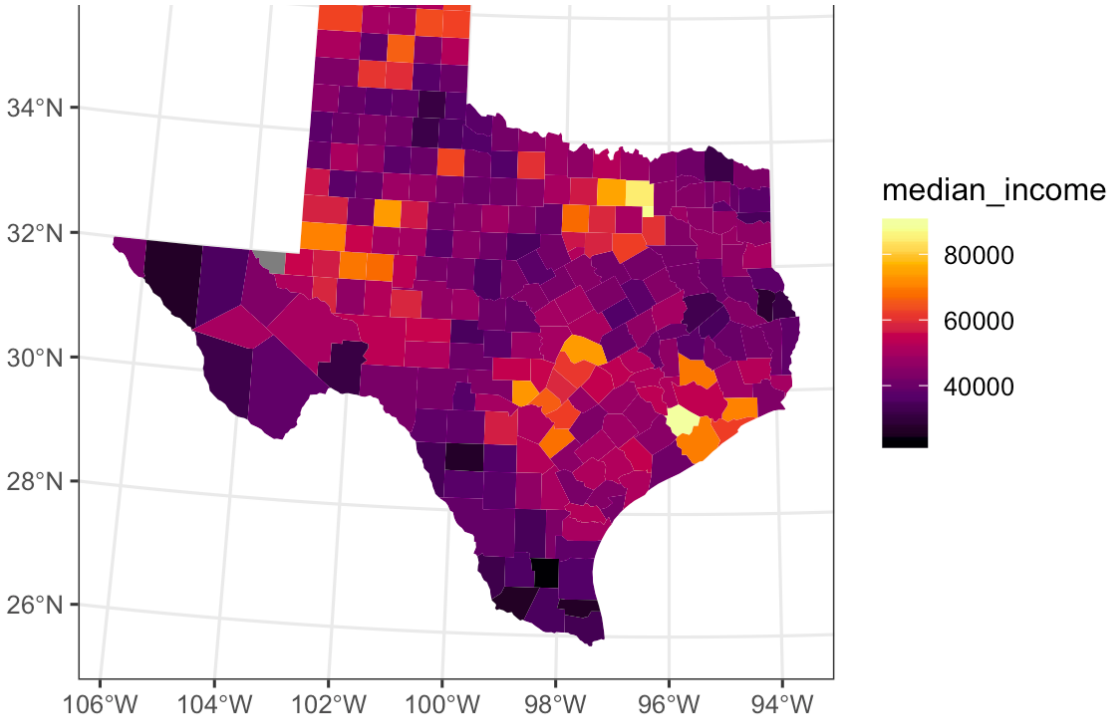
Using the data frame `US_counties_income`, plot all US counties, coloring each one by median income. Hint: Use `scale_fill_viridis_c(option = "B")` to create an appealing color effect.

```
ggplot(US_counties_income, aes(fill = median_income)) +
  geom_sf(color = NA) + # set `color = NA` to hide county boundaries
  scale_fill_viridis_c(option = "B")
```



Now plot only the counties of Texas, coloring each one by median income. Hint: Texas is represented by a code of `"48"`, and this code is stored in the `STATEFP` column in the data frame with county information.

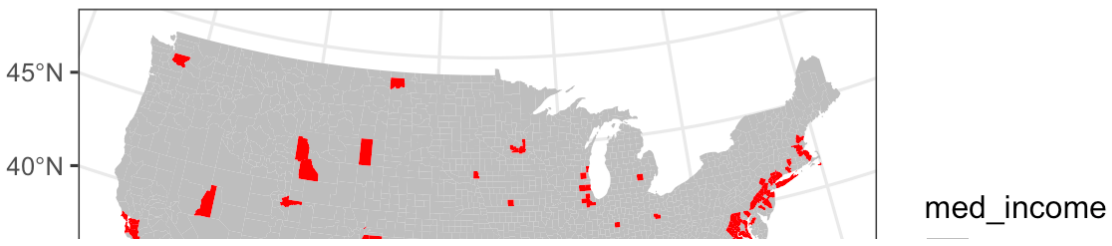
```
US_counties_income %>%
  filter(STATEFP == "48") %>%
  ggplot(aes(fill = median_income)) +
  geom_sf(color = NA) +
  scale_fill_viridis_c(option = "B")
```

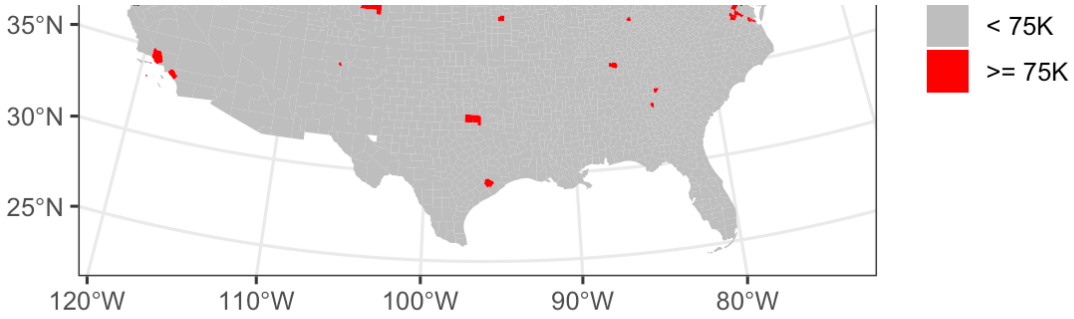


3. If this was easy

Make a map of all the counties in the lower 48, with counties with a median income of at least $75,000 highlighted in red.

```
US_counties_income %>%
  filter(!STATEFP %in% c("02", "15")) %>% # remove Alaska and Hawaii
  filter(!is.na(median_income)) %>% # remove counties with missing data
  mutate( # classify counties by high/low median income
    med_income = ifelse(median_income >= 75000, ">= 75K", "< 75K")
  ) %>%
  ggplot(aes(fill = med_income)) +
  geom_sf(color = NA) +
  scale_fill_manual(values = c("grey75", "red"))
```




35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
< 75K
>= 75K