Lab 2. Descriptive Statistics and Basic Mapping

GIS 3 - Geocomputation - Spring 2020 - Lily Cao

Contents

Load libraries

```
suppressMessages(library(sf))
suppressMessages(library(raster))
suppressMessages(library(dplyr))
suppressMessages(library(stringr))
suppressMessages(library(tidyr))
suppressMessages(library(spData))
suppressMessages(library(tmap))
```

```
## Warning: replacing previous import 'sf::st_make_valid' by
## 'lwgeom::st_make_valid' when loading 'tmap'
```

```
suppressMessages(library(maptools))
suppressMessages(library(cartogram))
suppressMessages(library(rgdal))
```

Choose a spatial dataset and load as a spatial data frame

```
df <- data.frame(us_states)
head(df)
```

```
##   GEOID        NAME   REGION             AREA total_pop_10 total_pop_15
## 1    01     Alabama    South 133709.27 [km^2]      4712651      4830620
## 2    04     Arizona     West 295281.25 [km^2]      6246816      6641928
## 3    08    Colorado     West 269573.06 [km^2]      4887061      5278906
## 4    09 Connecticut Norteast  12976.59 [km^2]      3545837      3593222
## 5    12     Florida    South 151052.01 [km^2]     18511620     19645772
## 6    13     Georgia    South 152725.21 [km^2]      9468815     10006693
```

```
##                         geometry
## 1 MULTIPOLYGON (((-88.20006 3...
## 2 MULTIPOLYGON (((-114.7196 3...
## 3 MULTIPOLYGON (((-109.0501 4...
## 4 MULTIPOLYGON (((-73.48731 4...
## 5 MULTIPOLYGON (((-81.81169 2...
## 6 MULTIPOLYGON (((-85.60516 3...
```

Description: I chose us_states, a spatial dataset from spData that contains data from the US Census Bureau and American Community Survey (ACS). There are 49 objects with 7 variables:
-GEOID: character vector of geographic identifiers.

-NAME: character vector of state names.

-REGION: character vector of region names.

-AREA: area in square kilometers of units class.

-total_pop_10: numerical vector of total population in 2010.

-total_pop_15: numerical vector of total population in 2015.

-geometry sfc_MULTIPOLYGON.

Source: https://www.rdocumentation.org/packages/spData/versions/0.3.3/topics/us_states

Provide summary statistics for key variables

```
summary(df)
```

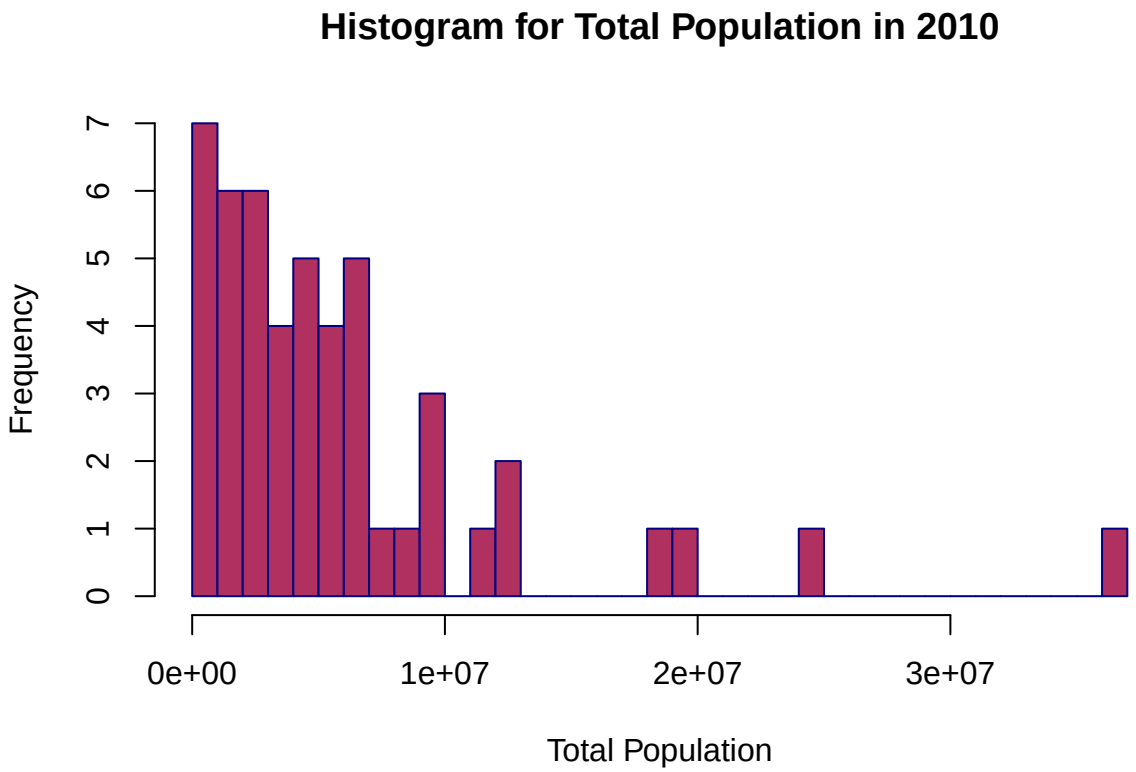
```
##     GEOID              NAME               REGION        AREA
##  Length:49          Length:49          Norteast: 9   Min.   :   178.2
##  Class :character   Class :character   Midwest :12   1st Qu.: 93648.4
##  Mode  :character   Mode  :character   South   :17   Median :144954.4
##                                        West    :11   Mean   :159327.3
##                                                      3rd Qu.:213037.1
##                                                      Max.   :687714.3
##   total_pop_10       total_pop_15              geometry
##  Min.   :  545579   Min.   :  579679   MULTIPOLYGON :49
##  1st Qu.: 1840802   1st Qu.: 1869365   epsg:4269    : 0
##  Median : 4429940   Median : 4625253   +proj=long...: 0
##  Mean   : 6162051   Mean   : 6415823
##  3rd Qu.: 6561297   3rd Qu.: 6985464
##  Max.   :36637290   Max.   :38421464
```

Description: The summary() function provides summary statistics for each of the 7 variables. For example, there are 9 rows (states) labeled "Northeast", 12 for "Midwest", 17 for "South", and 11 for "West" under the "REGION" column. For "AREA", "total_pop_10," and "total_pop_15", we're given the quartiles, min/max, and mean.

Generate a non-spatial plot of the variable of interest

```
hist(df$total_pop_10,
     main = "Histogram for Total Population in 2010",
```

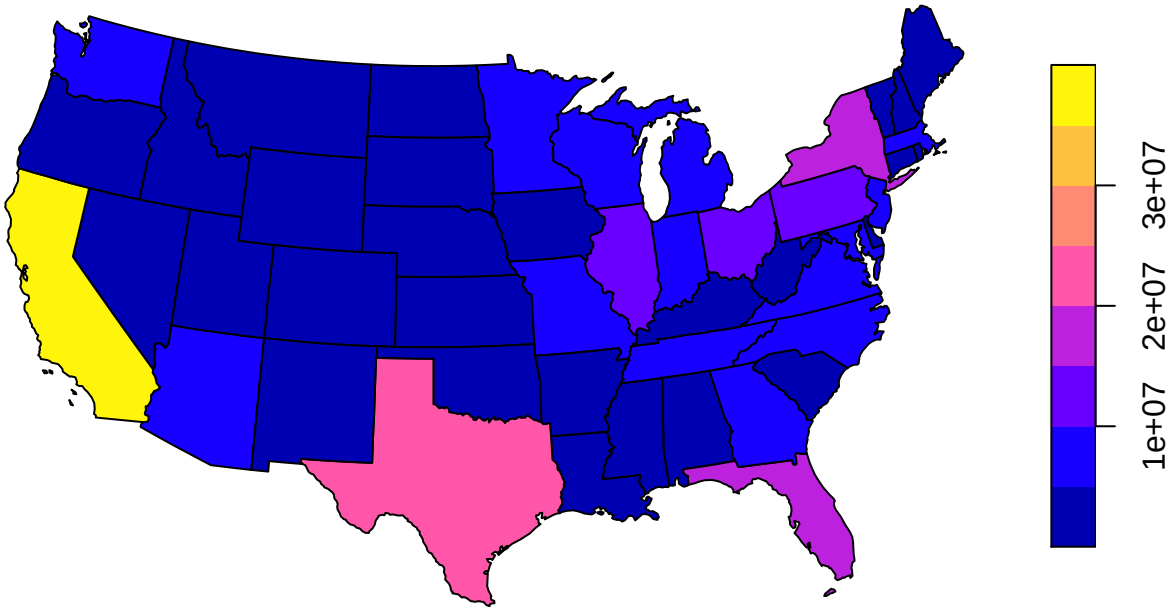
```
     xlab = "Total Population",
     border = "navy",
     col = "maroon",
     breaks = 50)
```



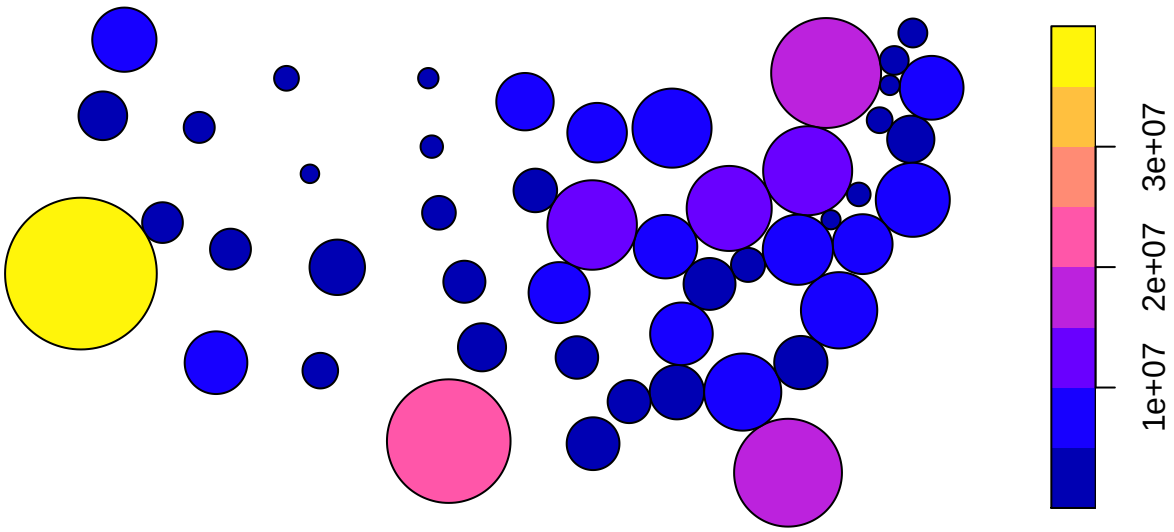
Description: The variable of interest here is total_pop_10 (the total population in 2010). The histogram is right-skewed, telling us that most states had populations less than 10,000,000 in 2010. The summary before told us that the maximum value for total_pop_10 is 36,637,290, which we can spot on the right end of the histogram (frequency = 1 so only one state has this value). I set the number of breaks (bins) to 50 so that the groups of total population are finer than the deafult.

Map the variable of interest

```
us_2163 <- st_transform(us_states, 2163)
plot(us_2163['total_pop_10'],
     main = "Map for Total Population in 2010")
```




```
us_dorling <- cartogram_dorling(us_2163, "total_pop_10")
plot(us_dorling['total_pop_10'],
     main = "Dorling Cartogram for Total Population in 2010")
```



Description: Before mapping total_pop_10, I re-projected us_states to equal area projections (US National Atlas Equal Area). By mapping total_pop_10, we can see which states and regions have the lowest or highest populations across the U.S. For example, California is obviously the state with the highest population in 2010, and many states from the West and Midwest have the lowest 2010 population. Using the "cartogram"

library and its cartogram_dorling function, I also created a Dorling cartogram, which uses sized circles to represent total_pop_10; the bigger the circle, the greater the population. This is useful because it's hard to differentiate the 2010 population differences for states with similar colors.