**R Module 4: Descriptive Spatial Statistics**

Tabular and visual explorations of data are always an important first step to understanding the distribution of actual values across a range of possible outcomes. Frequency tables, bar plots, histograms, and box plots are basic tools for this tasks. When we have spatially located data, mapping is an additional crucial tool.

These types of explorations have limits and we will also need actual measures of distribution of values across a range of possible outcomes and across geographic space. For this we rely on measures of central tendency and measures of dispersion.

Let’s start again with a shapefile of

**Importing and Visualizing Qualitative Data**

First, load the package to allow access to new commands in GISTools:

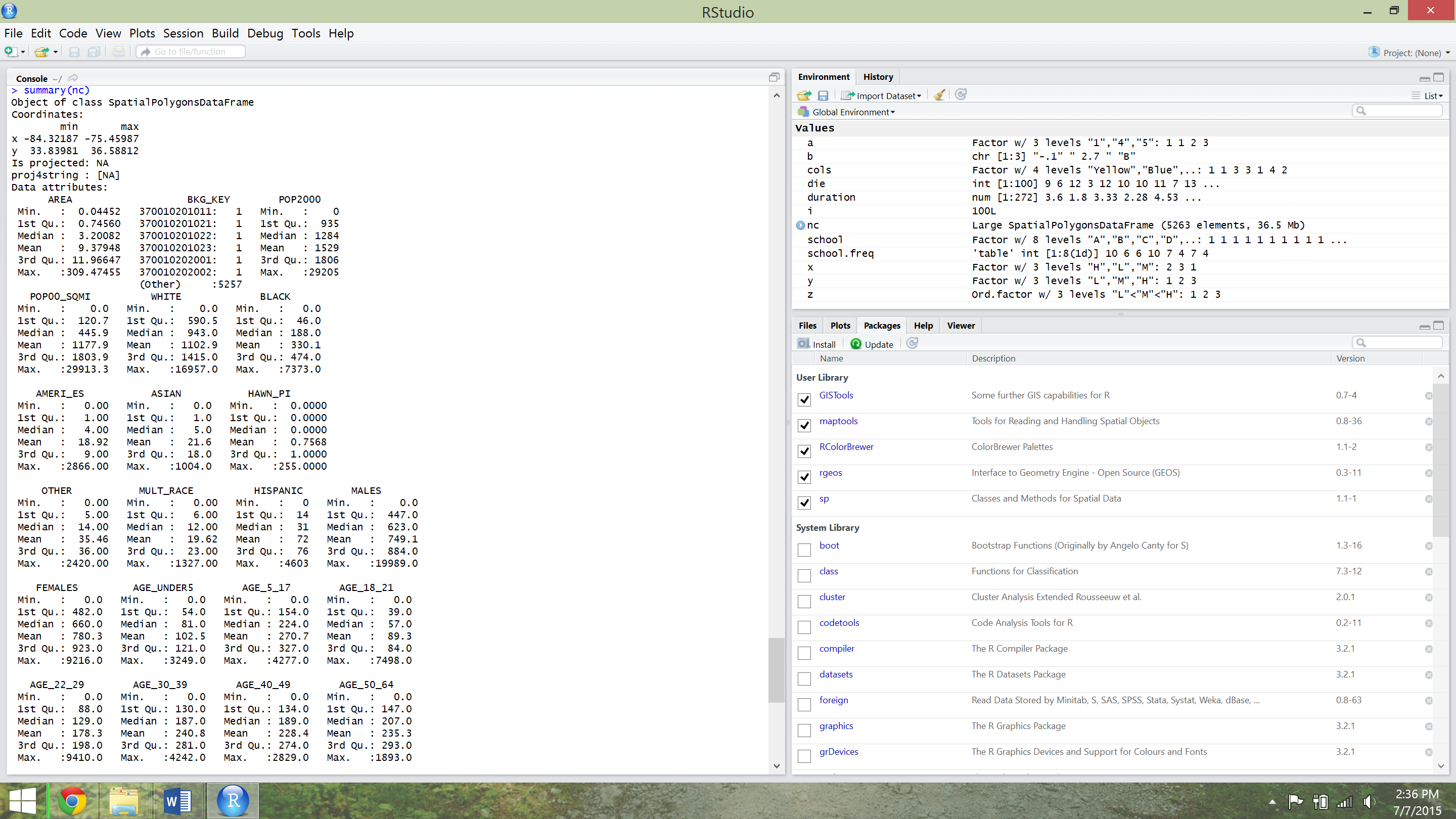
library(GISTools)

Next, let's read in an ESRI-style shapefile of census blocks in the state of North Carolina. You can download the zipped shapefile on the course website (unzip it before procedeing). I’ve included my pathway as an example – of course replace my pathway with your own when you type in the code.

nc=readShapePoly("C:/Users/Maggie/Downloads/ncblkgrp/ncblkgrp.shp")

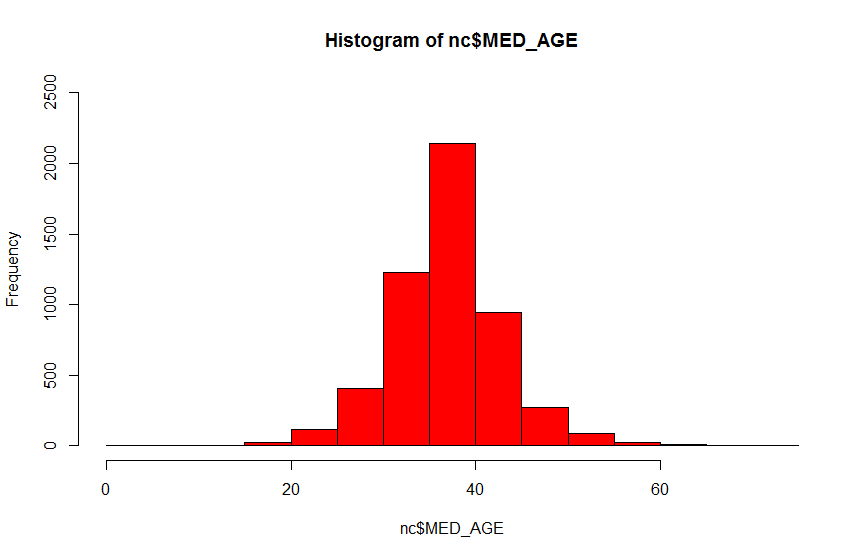
(Hint: To find you current workspace directory use the function “getwd()”, you can also set the work directory using the setwd() function).

Now let’s check out what variables are associated with the shapefile’s attribute table. summary(nc) You should see something like this:.



Just like last week, you can explore the frequencies and distributions of the quantitative variables using the hist() and boxplot() functions.

For example: hist(nc$MED\_AGE, col='red', ylim=c(0,80))



Of course, geographers are interested in spatial patterns so let’s explore this for the 2010 data using a simple choropleth map. choropleth(nc, nc$ MED\_AGE)



That’s quick and easy but not terribly helpful. We don’t know how the data is classified or what the breaks are. Keeping in mind that an effective visualization is a tool of communication, we’ll need to add some additional information to the map to reach this basic standard, such as a title and a legend.

So let’s use some more GISTools to increase our control over the map:

#let’s create our own classification.

#I’m setting the number of classes to 4 and changing the color palate to greens. shades=auto.shading(nc$MED\_AGE, n=4, cols=brewer.pal(4, "Greens"))

#let’s also add a legend

#the first two arguments set the location of the legend using the coordinate system of the map

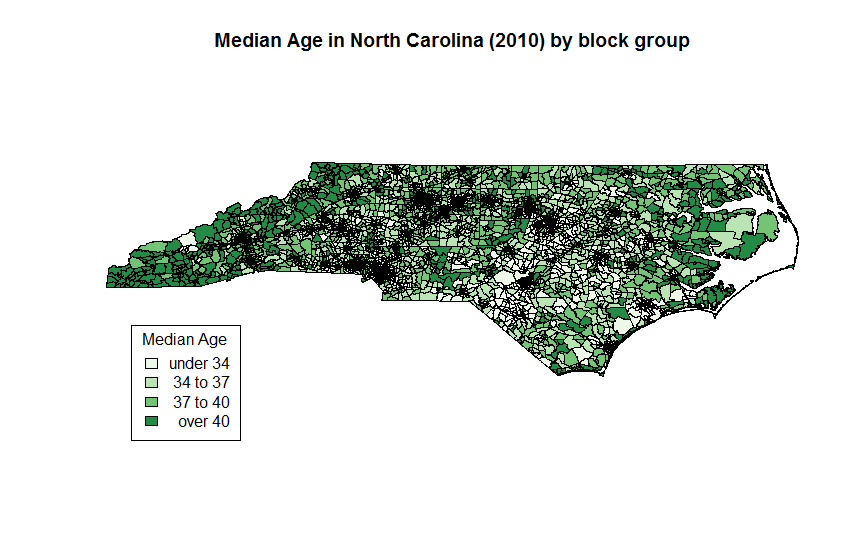
#then I link the legend to the number of classes and their respective colors before adding a title

choropleth(nc, nc$ MED\_AGE, shades)

choro.legend(-84, 34.5, sh=shades, title="Median Age")

#now add a title

title("Median Age in North Carolina (2010) by block group")



Better! Now we see the data classes (4) and the breaks between them (assigned by a default algorithm in the auto-shading function). Notice that we’re classifying data much like we did for histograms. But we also observed that small numbers of classes can mask variation in the data. Are we masking variation in the map?

**Explore this by changing the number of data classes and remaking the map. Provide the output/answers as requested below.**

**1.) Adjust the code to make a map of Median Age with 4, 6, and 8 classes. Provide each map with a legend and title.**

**2) How do these maps alter possible interpretations of the distribution of age across the state?**

Mapped data can lead to specific visual interpretations based on the number of data classes and the break points of the data classes used in the map. What if for example, if I’m trying to emphasize the prominence of the younger demographics on the map, I’ll probably use a different number of classes (say 4).

#I’m in charge of the map, not the other way round

#so let me prove my dominance by manipulating the output! Fun!

#I’m going to set the break points manually using the shading function

#first I create a new variable using shading

shades2=shading(c(20,30,40), cols=brewer.pal(4, "Reds"))

#now I’ll draw my new map

choropleth(nc, nc$ MED\_AGE, shades2)

#and redo the legend

choro.legend(-84, 34.5, sh=shades2, title="Median Age")

#now add a title

title("Median Age in North Carolina (2010) by block group")

Is there some middle ground between a black box approach (default in auto.shading) and conscious manipulation (deciding what to emphasize and setting breaks to produce that effect)? Of course there is! In auto-shading, you can specify either an equal interval approach (same numeric range in all classes) or quantile approach (equal number of observations in each class). But the overall number of classes is still up to you.

#let’s redo the map using the equal interval approach

#I have to use auto.shading again and specify a new argument called cutter

shades=auto.shading(nc$MED\_AGE, n=5, cutter=rangeCuts, col=brewer.pal(5, "Greens"))

choropleth(nc, nc$ MED\_AGE, shades)

choro.legend(-84, 34.5, sh=shades, title="Median Age")

#now let’s change to the equal observations in each class approach

shades=auto.shading(nc$ MED\_AGE, n=5, cutter=quantileCuts, col=brewer.pal(5, "Greens")) choropleth(nc, nc$ MED\_AGE, shades)

choro.legend(-84, 34.5, sh=shades, title=" Median Age ")

Lastly, for those of you who are more inspired, check out the following tutorial to add north arrows, scale bars, more shading options. We will present our maps in class and vote for the best one.

<http://rstudio-pubs-static.s3.amazonaws.com/958_d3123f6a9f95436a8177dd096ad768a7.html>

**In your report, choose a shapefile to map and analyze. Provide the following:**

**3.) A new map with title and legend with 5 data classes using rangeCuts (i.e. equal interval)**

**4.) Do it again for the same number of classes but with quantileCuts (i.e. equal number of observations in each class) instead**

**5.) Describe the difference between these maps. How are these maps likely to be interpreted differently (if at all all) using these different data classification methods?**

Note that all the maps you’ve made are of the exact same variable! And yet each map can lead to differing interpretations about how the observations are distributed across the state. This is why geographers place a high premium on data exploration – if you didn’t explore the data visually, you might never know just how sensitive your (or someone else’s) interpretations might be to the basic choices involved with data classification**.**

**To reinforce your ability to do this kind of exploration in R, produce the following for the optimal map for your shapefile (i.e. you can choose different break technique, different number of data classes, etc.).**

**6.) Display this map. Explain the number of data classes and break technique. Use histograms, descriptive statistics, or box plots to explain your choice**

**7.) Explain why this map is ideal. What are you trying to display?**