Bringing Geographical Data to Life: Usage of Chloropleth Maps

Justin Nelson

11/30/2017

Introduction

In class we've learned a variety of useful ways to interpret our data through more traditional visualizations such as histograms, line graphs, and barplots. However, one useful data visualization technique that we only skimmed the surface of in lab11 was the plotting of data onto geographical maps. Often times you'll run across datasets that will give statistics as they pertain to specific geographical regions. While you may able to conduct analysis on each individual region and then try to make sense of the results within a written table, it is often helpful to visualize the analysis results on a familiar geographical map. This is where we can utilize the combination of the maps and ggplot2 package to our benefit.

Motivation

I feel fortunate to have receieved my high school education from a public highschool that was well regarded and put noticable effort into pushing its students towards success, and I wish all young adults in this nation could receive the same quality of free education. However, there were many high schools just several miles away that had students who struggled to stay motivated to graduate, strive for solid academic achievement, or attend a university. I believe that a potential reason for this disconnect may be due to the funding a high school recieves from the state. To help determine whether or not this may be a cause for the separate outcomes, we'll be analyzing if there is a correlation between the funding a high school receives per student relevant to percentage of high school students who take the SAT and their performance on the SAT itself. Looking at the percentage who take the SAT and the associated perforamnce can serve as a good indicator of students' attitude towards education, as the SAT is a commonly used tool to achieve higher education.

Background

We've used ggplot2 extensively in Stat 133 to plot a variety of data. maps lets us take the traditional ggplot2 package one step further. As we know, we use the '+' sign when using the ggplot2 package to add aditional layers to the visualization. For this activity, we'll be adding an additional layer to a map create a map known as a "choropleth map". A choropleth map is a, "thematic map in which areas are shaded or patterned in proportion to the measurement of the statistcal variable being displayed on the map." maps allows us to add another layer to our visualization with the same '+' operator, but the way in which we specificy which map to use and how to color it is a bit tricky. Let's take a high level look at the steps involved before diving in:

1. Calculate appropriate statistics for each region in which you are interested
2. Define and specifiy a desired region, and call ggplot on it to generate a blank map
3. Add second layer to map that overlays appropriate statistics with varying levels of fill colors to denote differences

With that outlined, let's hop into the example and analysis.

Example

First, let's acquire and load in the required libraries for our analysis. We'll be using dplyr to aid in the data processing and ggplot2 and ggmaps for the actual visualization.

```
# Load all required packages
library('dplyr')
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library('ggplot2')
library('maps')
```

With the libraries loaded in, let's now grab our data.

The dataset we're using can be found at this github link: https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/car/States.csv

The CSV contains 51 rows (one for each state), and 8 columns.

```
# Save the web address as a variable
site <- 'https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/car/States.csv'

#Download the contents of the file to a local directory, and then use the read.csv function to acquire the data, as it's stored as a CSV file online and on our machine.
download.file(site, destfile = 'data/state_education_data.csv')

dat <- read.csv('data/state_education_data.csv')
```

The columns we'll use are the state (X), the SAT verbal score (SATV), SAT math score (SATM), percentage of students who take the SAT (percent), and state spending per student in thousands of dollars (dollars). We'll also sum together the two different SAT scores in each state in order to get an aggregate, which will hopefully give us a more well-rounded view of students' performance.

```
dat <- dat %>%
  rename(state = X) %>%
  select(state, SATV, SATM, percent, dollars) %>%
  mutate(SAT = SATV + SATM)
head(dat)
```

```
##   state SATV SATM percent dollars SAT
## 1    AL  470  514       8   3.648 984
## 2    AK  438  476      42   7.887 914
## 3    AZ  445  497      25   4.231 942
## 4    AR  470  511       6   3.334 981
## 5    CA  419  484      45   4.826 903
## 6    CO  456  513      28   4.809 969
```

Before jumping into mapping these data points onto a every individual state and observing any differences, it may be worth doing some prelimanary analysis on the entire country. Let's take a look at the correlations between the variables that we talked about at the beginning of this post (As a reminder, spending on student compared to percent of students who take the SAT, and spending on student compared to SAT cumulative score).

```
# Correlation between spending on student and percentage of students who take SAT
percent_cors <- format(cor(dat$dollars, dat$percent), digits = 2)
print(percent_cors)
```

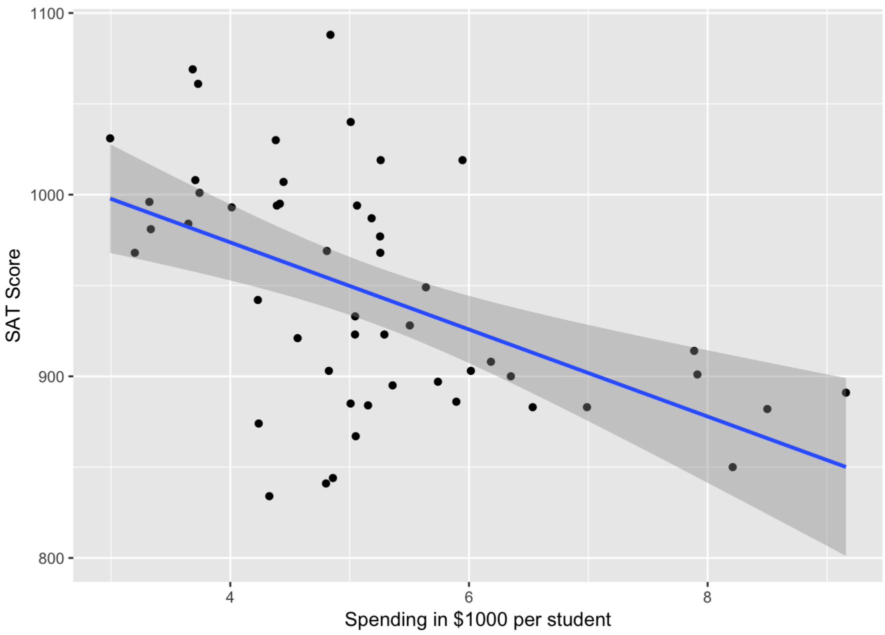
```
## [1] "0.71"
```

```
# Correlation between spending on student and scoring on SAT
score_cors <- format(cor(dat$dollars, dat$SAT), digits = 2)
print(score_cors)
```

```
## [1] "-0.51"
```

As one may expect, there is a significant positive correlation between an increase in state spending per student and the completion of the SAT. However, something that comes as a surprise is the negative correlation between spending and performance on the test. As spending increases, performance tends to go down. Let's take a look at a plot of this:

```
ggplot(dat, aes(x=dollars, y=SAT)) + geom_point() + labs(x='Spending in $1000 per student', y = 'SAT Score') + geom_smooth(method = 'lm')
```



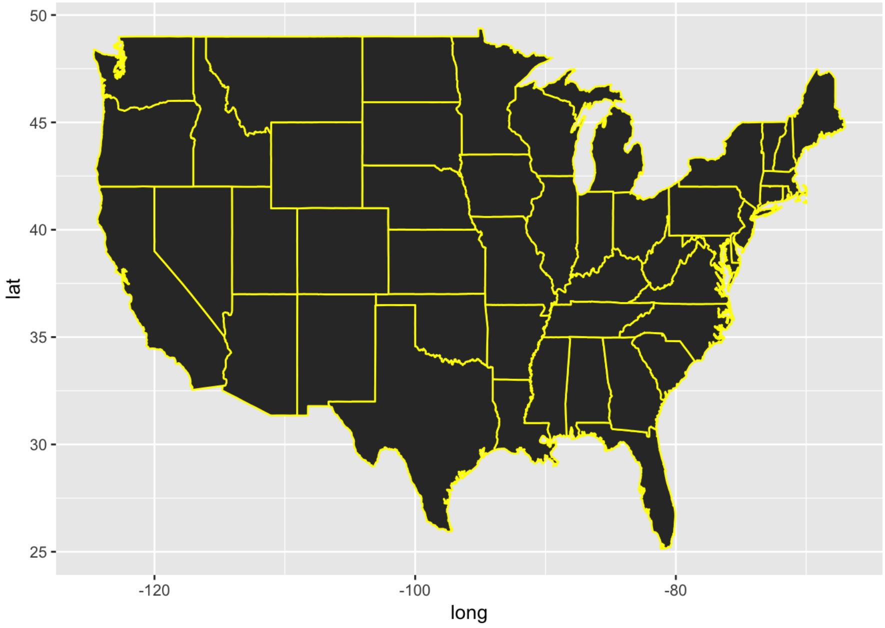
One reason that comes to mind as to why this may be is that if a state has a high rate of students who are struggling on standardized tests, they may pump more money into their education system, whereas a state who has high performing children will not necessarily need to spend as much. This is just a simple hypothesis, and

Without being able to tell much more from that simple scatter plot, let’s look at a map of the entire United States in hopes to uncover some regional differences that we may be able to make more inferences from.

To begin the mapping process, we first need a general outline of the U.S. to be generated. maps has a built-in function called map_data that takes an argument of a some type of geographic region, and returns pre-loaded data. For our purposes, we can pass in ‘state’, and map_data will return a dataframe that contains the geographical coordinates for the outlines of all 50 states.

Once we have the outlines of the states, we can use a combination of ggplot and geom_polygon on the dataset to generate a basic map.

```
#Acquiring outlines of all 50 states with map_data
states <- map_data('state')

# Generating a map of the United States using the latitude and longitude coordinates provided from map_data.

us <- ggplot(states, aes(x = long, y = lat)) + geom_polygon(aes(group = group), color = 'yellow')

us
```



We now have two data frames - one with information on educational attainment, and one with data on geographical locations. Both of them have a common variable, state, so we’ll join the tables by that.

(Note we’ll have to a do a little string manipulation first in order to turn full state names into abbreviations).

```
library('Hmisc')
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
# String manipulation to be abel to join tables
dat <- dat %>%
  mutate(state = state.name[match(dat$state, state.abb)])

states <- states %>%
  mutate(state = capitalize(region))
# Using inner_join to join the two tables by the 'state' variable
joined <- inner_join(dat, states, 'state')
```

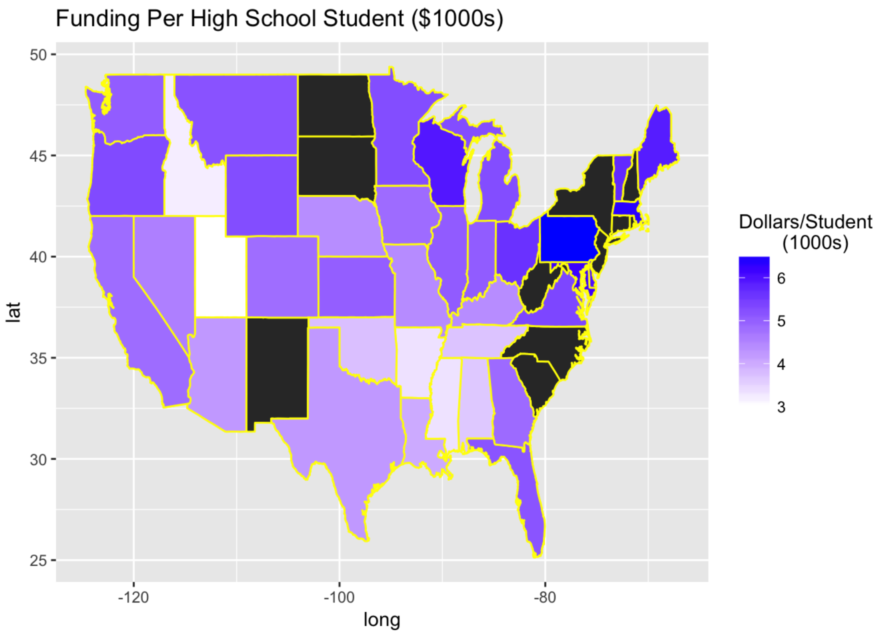
With the tables now joined, we can generate the map. We’ll make maps that display funding in each state, percentage of students who take the

SAT in each state, and performance on the SAT in each state. To make each map represent the variable we're interested in, we'll set the 'fill' argument within the aesthetics parameter to the column of data we want to look at on our map.

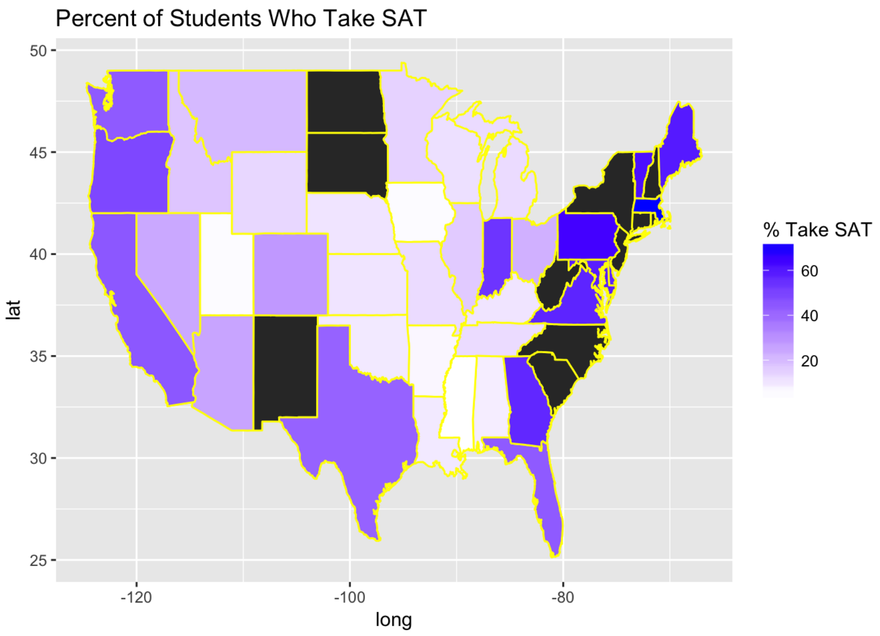
(Notice that there are several black states. These are states in which there was no data aviailable.)

```
# Use our base map of the United States generated above, and then add another
# layer to it using geom_polygon. Use our 'joined' data frame as the data to be used,
# set 'fill' to the variable you want to analyze, and and then use the 'scale_fill_gradient' # function to pick your colors.
funding_map <- us + geom_polygon(data = joined, aes(x = long, y = lat ,group = group, fill = dollars), color = "yellow") + scale_fill_gradient(low = "white", high = "blue", name = 'Dollars/Student
        (1000s)') + ggtitle('Funding Per High School Student ($1000s)')

# Apply this same mapping process to the other two variables.

take_map <- us + geom_polygon(data = joined, aes(x = long, y = lat, group = group, fill = percent), color = "yellow") + scale_fill_gradient(low = "white", high = "blue", name = '% Take SAT') + ggtitle('Percent of Students Who Take SAT')

score_map <- us + geom_polygon(data = joined, aes(x = long, y = lat, group = group, fill = SAT), color = "yellow") + scale_fill_gradient(low = "white", high = "blue", name = 'SAT Score') + ggtitle('SAT Scores')
```

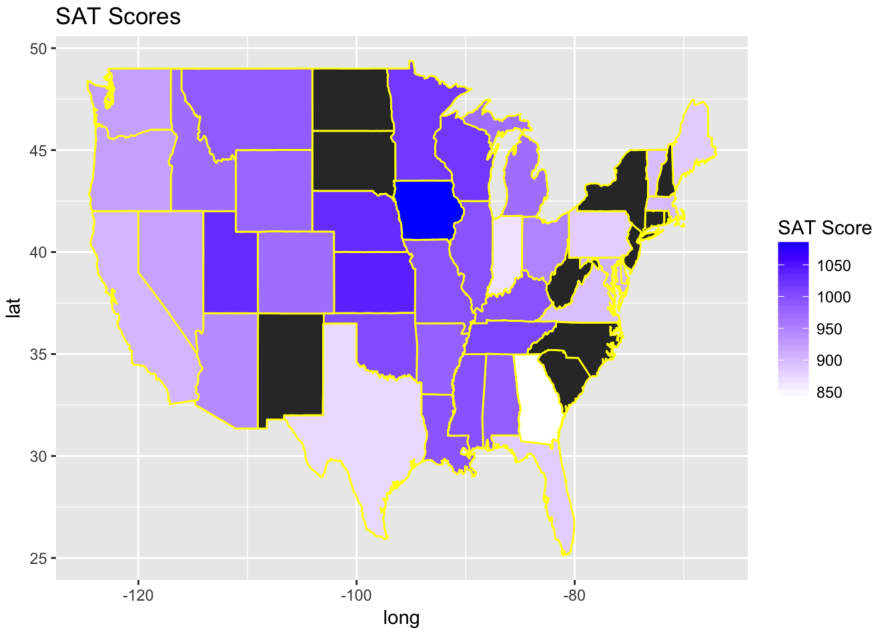
```
funding_map
```


Funding Per High School Student ($1000s)

```
take_map
```


Percent of Students Who Take SAT

```
score_map
```



Discussion

There are some interesting takeaways from the above visualizations, primarily in the midwest and southern United States. It appears that those states within the region receive a relatively normal or even high amount of funding, but the percent of students who take the SAT is low compared to the entire nation. However, of those students who do take the SAT in those regions, they tend to perform better than average.

Obviously these are some very general conclusions, and one would need to dig much deeper to find out why this may be the case. For example, what type of socioeconomic factors are at play in this region? What is the high school graduation rate in each state?

Conclusion and Take Home Message

I'm impressed with the mapping abilities that R provides, especially with packages that we have some familiarity with from this class. While the main driver to producing the actual visualization is ggplot2, the maps package is integral to supplying the data required to make the visualization.

Chloropleth maps are generally intuitive to read for most people, and knowing how to create one in R should prove useful if you ever have to visualize data over a geographic region. Visualizations of data such as this one are powerful in that it allows one to present a large amount of data in a quick manner, and can also help prove a point better than just displaying someone numbers.

Remember that you're not just limited to generating chloropleth maps of the United States - you can apply this same methodology to any geographic region! I hope that you found this post useful and are able to utilize it in your life and or career at some point.

References

Chlorpleth Map - https://en.wikipedia.org/wiki/Choropleth_map

Making Maps in R - https://cengel.github.io/rspatial/4_Mapping.nb.html

How to Make Chloropleth Maps with R - http://bl.ocks.org/prabhasp/raw/5030005/

Geographic Visualizations with R's ggmaps -https://blog.dominodatalab.com/geographic-visualization-with-rs-ggmaps/

ggmaps package - https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Education Data - https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/car/States.csv

Mapping in R - http://eriqande.github.io/rep-res-web/lectures/making-maps-with-R.html