Post02-Aiden-Rafii

Post 02 - Using GGgplot2 to make choropleths of education and crime data in the United States

Introduction & Background

Choropleth maps, “display divided geographical areas or regions that are colored, shaded or patterned in relation to a data variable. This provides a way to visualize values over a geographical area, which can show variation or patterns across the displayed location.” Being able to map statistics to locations on maps is a visual that is extremely useful for explaining your findings. It is both more interesting to the reader and easier to visualize. In this post I will show the reader how to make a map of the United States using ggplot2 and how to visualize different data sets on these maps. I will also show the reader how to graph a loess smooth line to view the relationship between two variables.

The data we use and manipulate is state data on education and violent crime. We will be checking out whether or not education spending, teacher salaries, and SAT scores have a direct link to the amount of violent crime.

Motivation

With all the talk that’s been going on about how the country should spend our tax money, I wanted to check if one area of spending, education, has a profound effect on the violent crime rate. This would no doubt make education a more valuable commodity. In this post I hope to visualize the effects of education on violent crime and view which states spend the most on education.

Packages & Content

In order to reproduce this post, you will need to load these 4 packages. If you do not have them you can use install.packages() to install them.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.4.2
```

```
library(maps)
```

```
## Warning: package 'maps' was built under R version 3.4.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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
```

Datasets

I went on Github to find these certain data sets. One pertains to the arrests in the USA by state and by crime. The other is a data frame of education spending, SAT math and verbal scores, and average teacher salary.

```
#Here I download the crime data and store it in a folder

URL <- "https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/datasets/USArrests.csv"
download.file(URL, destfile='crime_data.csv')
crimedat <- read.csv('crime_data.csv')

#I change the column name of state to region
colnames(crimedat)[1] <-"region"

#give every row its own unique ID number to merge with education data
crimedat$ID <- seq.int(nrow(crimedat))

#Here I download the education data and store it in a folder

URL <- "https://raw.github.com/vincentarelbundock/Rdatasets/master/csv/car/States.csv"
download.file(URL, destfile='education_spending.csv')
educdat <- read.csv('education_spending.csv')

#I remove row 9 which is the D.C
#our crime data does not have data for DC. DC also is not its own state.
educdat <- educdat[-c(9),]

#give every row its own unique ID number to merge with crime data
educdat$ID <- seq.int(nrow(educdat))

#Merge the two data frames by the ID numbers that I gave each state.
crime_and_education<-merge(educdat, crimedat, by="ID", all.x=TRUE)

#I change the column with State values to the names of all the states.
#I then change that column name to region in the next line
crime_and_education$X <- crime_and_education$region.y
colnames(crime_and_education)[2] <- "region"

#I drop extraneous columns I don't need
crime_and_education <- crime_and_education %>%
  select(-c(region.x, ID,UrbanPop, region.y))

#I add a column called total crime that is a combination of the other 3 violent crime variables
crime_and_education$totalcrime = crimedat$Murder +
  crimedat$Assault + crimedat$Rape
crime_and_education$region = as.character(crime_and_education$region)
```

Correlation between Educational Variables and Violent Crime

I'm very interested to see how educational spending, average teacher salary, and SAT scores affect violent crime. I believe that spending more on education reduces violent crime and that higher SAT scores and teacher salaries will lead to less crime as well.

```
#Total SAT scores
crime_and_education$SAT <- crime_and_education$SATM + crime_and_education$SATV
#3 correlations of educational variables and Total Crime

corr_dollars <- round(cor(crime_and_education$totalcrime,
                          crime_and_education$dollars),2)

corr_salary <- round(cor(crime_and_education$totalcrime,
                         crime_and_education$pay),2)

corr_SAT <- round(cor(crime_and_education$totalcrime,
                      crime_and_education$SAT),2)

print(c(corr_dollars, corr_salary, corr_SAT))
```

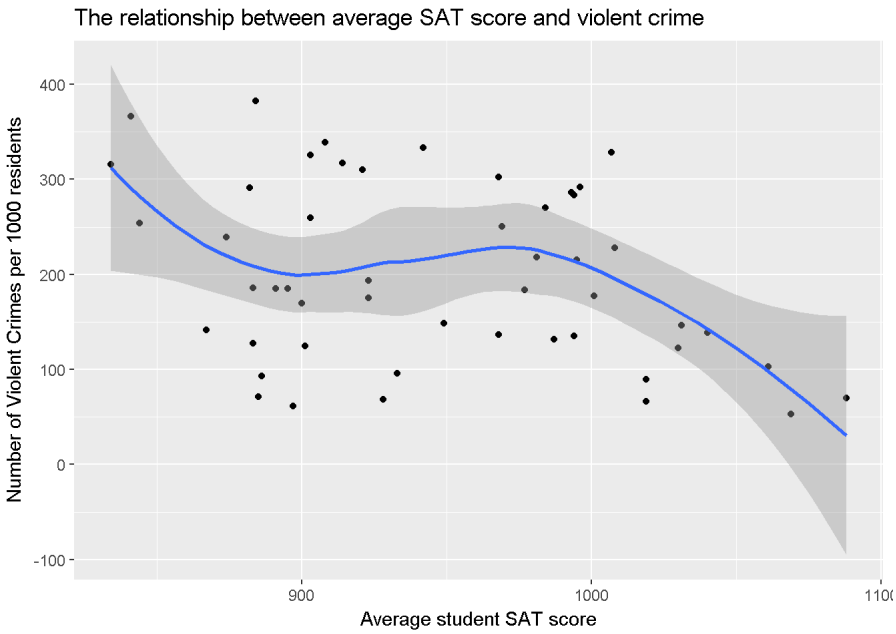
```
## [1] -0.02  0.21 -0.33
```

We can see from our correlation results that though most likely not statistically significant due to Omitted Variable Bias that education spending does not have much of an effect of total crime, teacher salary has a positive effect on crime and higher SAT scores have a negative effect on violent crime.

Graphing average SAT scores and violent crime

```
ggplot(crime_and_education, aes(x = crime_and_education$SAT,
                                y = crime_and_education$totalcrime)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Average student SAT score",
       y = "Number of Violent Crimes per 1000 residents",
       title = "The relationship between average SAT score and violent crime") +
  geom_text(x = 30, y = 50, label = paste("r =", corr_SAT))
```

```
## `geom_smooth()` using method = 'loess'
```



From our plot and correlation coefficient we can see that there is a slight negative relationship between average SAT score and violent crime.

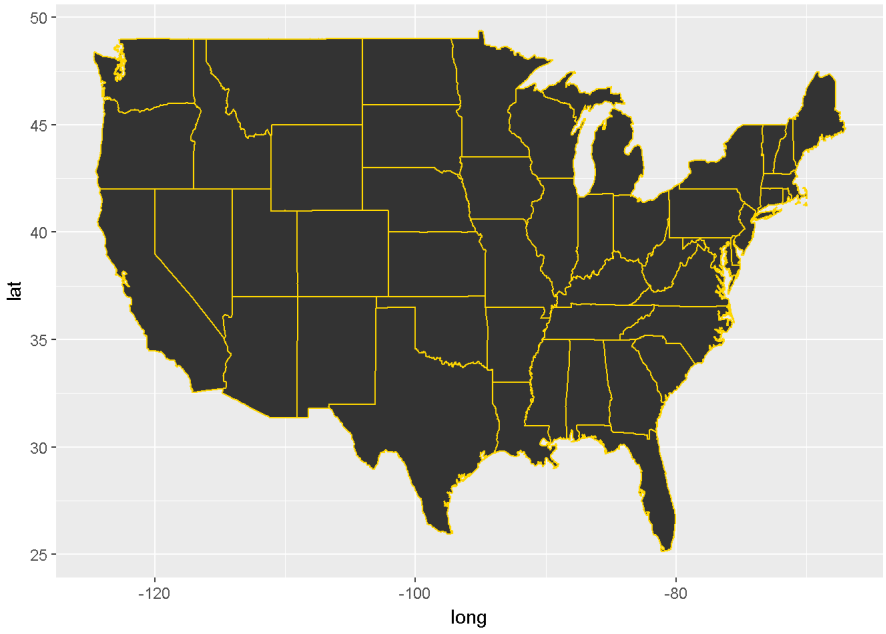
Mapping our data

We will use our state data set from the maps package we loaded to build a map of the United States in R. The states data set uses longitude and latitude points throughout a coordinate in order to build a map. We will use ggplot and geom_polygon to build this map. Geom_polygon will build a shape according to the coordinate points that you input which we have from the data set state.I change the color of the lines of the state boundaries using color = gold in my geom_polygon.

```
states <- map_data("state")
crime_and_education$region <-  tolower(crime_and_education$region)

map <- ggplot(data = states) +
  geom_polygon(aes(x = long, y = lat, group = group),color = 'gold', size = .5)

map
```



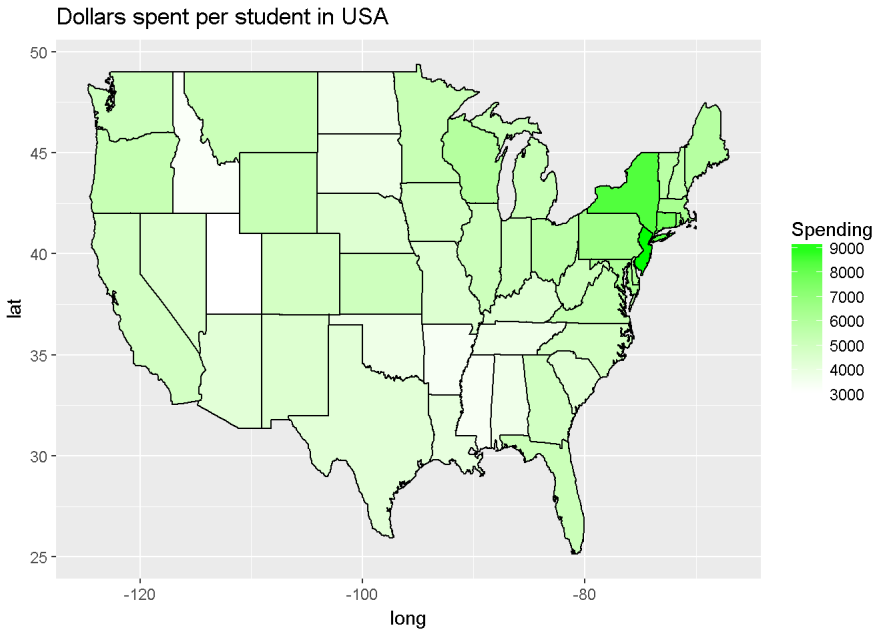
Now that we've built our basic map we're going to go ahead and merge our crime and education data with our mapping data. I merged them by region. Make sure to set fill in your geom_polygon portion of the function to the variable you'd like to map on your map of the United States. In this example I used dollars spent per student in the USA.

```
#Changing dollars spent per student in the 1000s to total dollars spent per student
if(max(crime_and_education$dollars) < 11){
  crime_and_education$dollars <- crime_and_education$dollars*1000
}else if (max(crime_and_education$dollars) > 10000){
  crime_and_education$dollars <- crime_and_education$dollars/1000
}

#joining the data of the states data frame and our education/crime merged data frame
crime_educ_maps <- inner_join(states, crime_and_education, by = 'region')


#Use labs to title my map
#use name in scale_fill_gradient to give  a name to my density bar on the right.
#use color in scale_fill_gradient to attribute colors to my map in high and low density areas..
dollarmap <- map + geom_polygon(data = crime_educ_maps,
                                aes(x = long, y = lat ,
                                    group = group,fill = dollars),
                                color = "black") +
                  labs(title= "Dollars spent per student in USA") +
                  scale_fill_gradient(name="Spending",low="white", high="green")

dollarmap
```

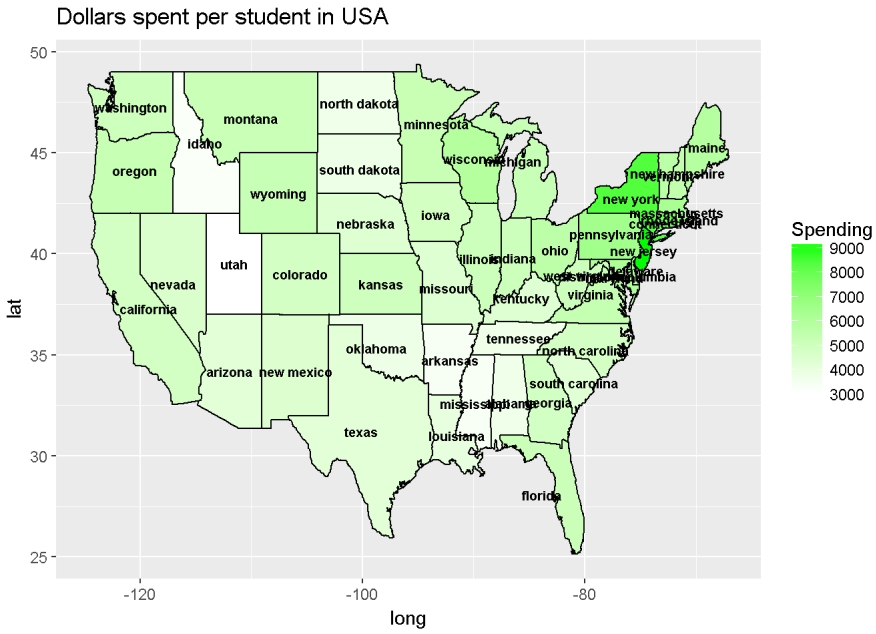


Label your states

Often times you will want to label your states. There are many different ways to do this. The way I like to do it is getting the average range of the longitude and latitude coordinates of each state. This usually will add the labels in an area in the middle of a state. We can then use geom_text to display those names on our map with the average coordinates we have provided.

```
#averaging the longitude and latitude ranges of the coordinates in each state
#then add them into a data frame with state names
state_names <- aggregate(cbind(long, lat) ~ region, data= states,
                         FUN=function(x) mean(range(x)))

#add labels with geom_text using the statenames data frame
#change size of text and font style with fontface and size
dollarmap2 <- dollarmap +geom_text(data=state_names,
                                   aes(long, lat, label = region),
                                   size=2.5, fontface = 'bold')
dollarmap2
```



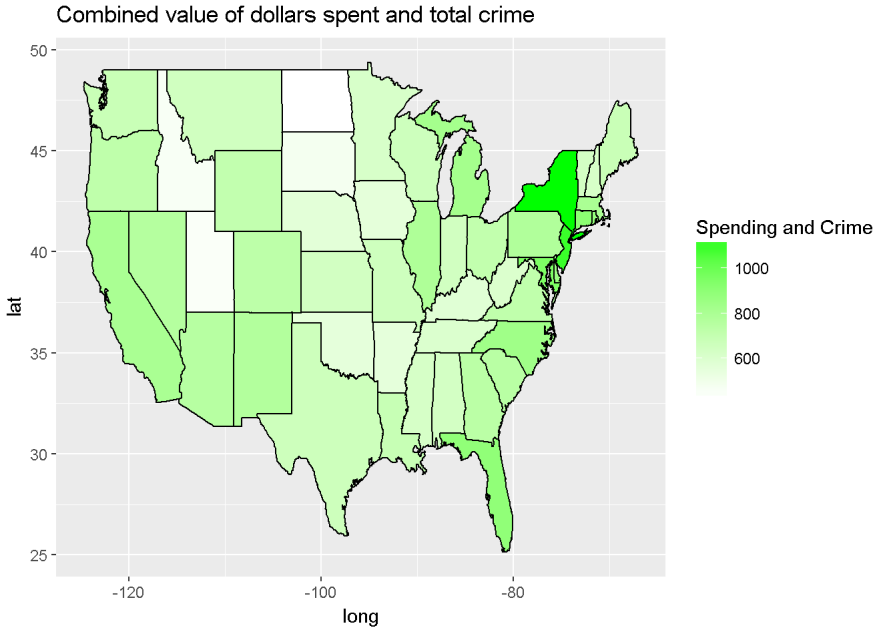
Crime and Spending map

In an attempt to combine the values of dollars spent per student and total crime I add the two values together and make one map using the combined values in order to see if there was any difference between the original map and this second one. Our correlation for the two variables was close to 0 so we should not see much of a difference.

```
#divide the dollar value by 10 to make sure it has the same number of digits as the total crime statistics
crime_educ_maps$dollarsandcrime <- crime_educ_maps$dollars/10 +
  crime_educ_maps$totalcrime

crime_and_money<- map + geom_polygon(data = crime_educ_maps,
                                     aes(x = long, y = lat ,group = group,fill =
                                           dollarsandcrime),
                                     color = "black") +
                        labs(title= "Combined value of dollars spent and total crime") +
                        scale_fill_gradient(name="Spending and Crime",low="white", high="green")

crime_and_money
```



There is not much of a difference in colors between the two maps.

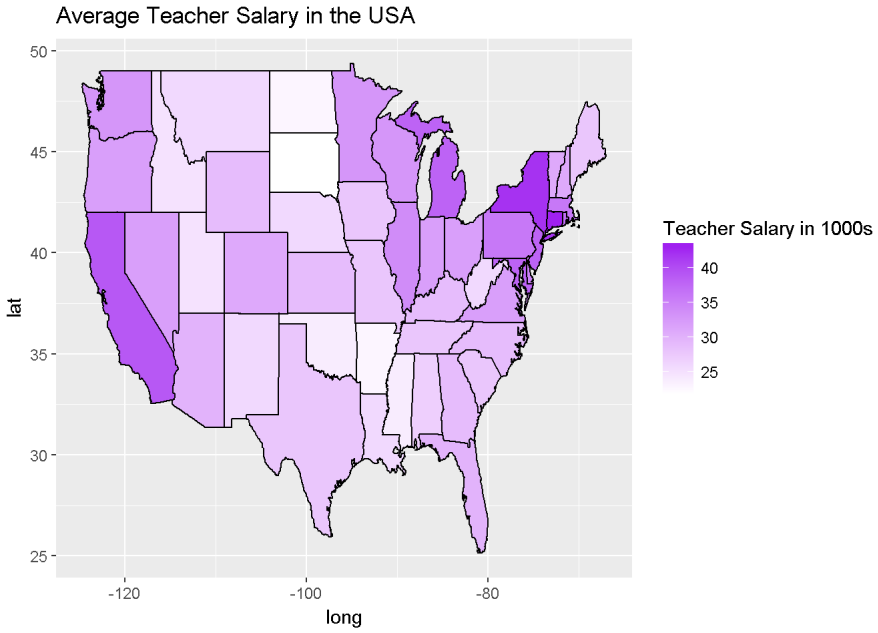
Mapping average teacher salary

Here we will use the same technique we used to map our spending map in order to map average teacher salary

```
#We use the same technique as making the dollar map
#We make our map purple this time by changing the fill gradient to purple

pay_map <-  map + geom_polygon(data = crime_educ_maps, aes(x = long, y = lat ,
                                                           group = group,fill = pay),
                               color = "black") +
                  labs(title= "Average Teacher Salary in the USA") +
                  scale_fill_gradient(name="Teacher Salary in 1000s",low="white", high="purple")

pay_map
```

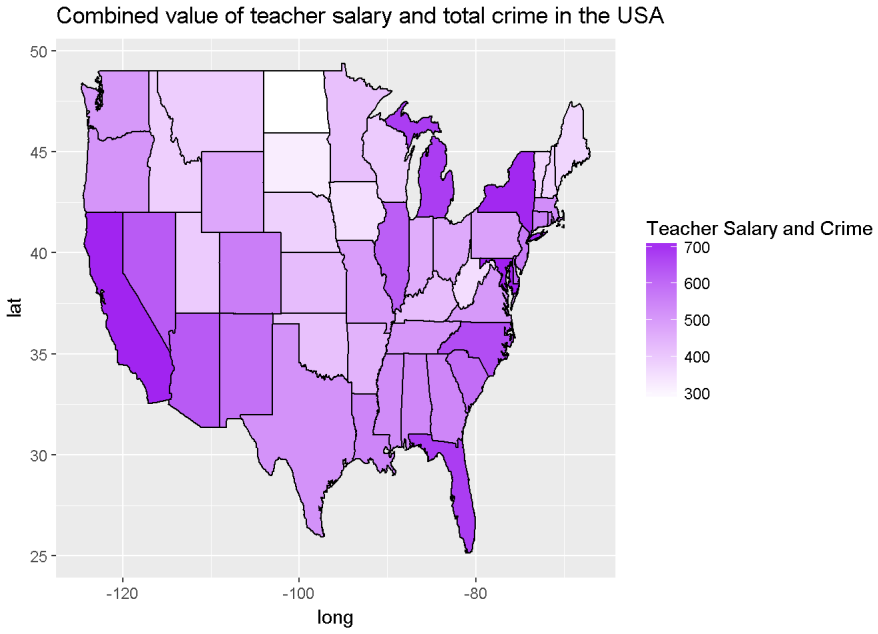


Crime and average teacher salary map

In an attempt to combine the values of average teacher salary and total crime I add the two values together and make one map using the combined values in order to see if there was any difference between the original map and this second one. Our correlation for the two variables was around .21 so we should see most of the states become a little bit darker compared to our map of just using average teacher salary since a higher teacher salary points to a higher crime rate in our data.

```
#multiply the salary value by 10 so it has the same number of digits as the total crime statistics
crime_educ_maps$payandcrime <- crime_educ_maps$pay*10 + crime_educ_maps$totalcrime

crime_and_money<- map + geom_polygon(data = crime_educ_maps,
                                     aes(x = long, y = lat ,
                                         group = group,fill = payandcrime),
                                     color = "black") +
                        labs(title= "Combined value of teacher salary and total crime in the USA") +
                        scale_fill_gradient(name="Teacher Salary and Crime",
                                            low="white", high="purple")
crime_and_money
```



This map definitely looks a lot more purple than the previous one insinuating that teacher salary does have a positive relationship with total crime.

This could be due to the idea that though the educational spending per student is relatively similar no matter the population of the state, smaller states with less population and less tax dollars will pay their teachers less. Therefore teachers that are paid more will be in states with a higher population which could attribute to the crime rate.

Saving our map

we can save our image using ggsave

```
ggsave(dollarmap2, file="dollarmap2.png")
```

```
## Saving 7 x 5 in image
```

Conclusion

While viewing our maps we can see that the northeast states dominate in education spending. New York appears to be the biggest spender in education per student. We can ask ourselves whether or not this is merely because they have more tax dollars due to a larger population, or if this is a genuine policy of New York's government to spend more on education.

Where it gets murky is when relating crime rates to educational variables. The problem with this data is that there are endless amounts of variables that lead to crime rate not just educational variables. This leads to omitted variable bias. However, from this we gain glimpses into indicators of a higher or lower crime rate.

The idea that higher SAT scores could lead to a lower crime rate is a very interesting idea. Maybe the SAT score is indicative of the type of support a student receives at home in the form of having a family willing to support the student and pay for tutoring.

Though there are tons of variables that go into crime rates and educational attainment, we can use maps and graphs with ggplot2 to help us visualize our data and bring to light very interesting relationships in our data.

References

1. Building choropleths in R - http://rforpublichealth.blogspot.com/2015/10/mapping-with-ggplot-create-nice.html
2. Tutorial on building choropleths and merging your data sets - https://trinkerrstuff.wordpress.com/2013/07/05/ggplot2-choropleth-of-supreme-court-decisions-an-tutorial/
3. Mapping spatial data using ggplot - https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/ggmap/ggmapCheatsheet.pdf
4. How to sequence your data frame rows in order to merge them -https://stackoverflow.com/questions/23518605/add-an-index-numeric-id-column-to-large-data-frame
5. Education and the effects on Incarceration - http://prospect.org/article/education-vs-incarceration
6. Mapping United States census with GGmap - https://www.youtube.com/watch?v=EtJ-iTZeqTg
7. choropleth basics - https://datavizcatalogue.com/methods/choropleth.html