Spending per Pupil Effects Incarceration Rates per State

Giovanni Rivera and Michael Minton

2023-10-18

```
#Begin work by loading the appropriate packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(skimr)
library(dplyr)
library(ggplot2)
library(usmap)
```

Title and Introduction The United States has the highest incarcerated population in the world with a total of 1.76 million in 2021 according to the World Prison Brief. Addressing the issue of mass incarceration in the United States has been a hot political topic for decades. Considering that estimates by the Prison Policy Initiative place the cost of operating federal, state and local correctional facilities at $265 billion dollars in 2012 alone.

Considering the significant impact incarceration has on many aspects of individuals daily lives, Giovanni and myself were interested in how educational quality and resources affect these rates.

Acknowledging and understanding the root causes of this phenomenon is essential if the United States hopes to address the current incarceration dilemma that is highly prevalent in American society.

For our research, we focused on data from the United States Census Bureau, 2021 Public Elementary-Secondary Education Finance Data, Data from the Bureau of Justice Statistics, Prisoners in 2021 – Statistical Tables, and data from the National Center of Education Statistics, Number of operating public schools and districts, student membership, teachers, and pupil/teacher ratio, by state or jurisdiction: School year 2019–20.

One note for our data is that the pupil/teacher ratio is drom 2019-2020 as there was not avialable data for that time due to Covid-19.

Each unique row in these datasets represent an individual state at a point in time. These datasets all contain states as a primary variable meaning joining these sets by state would be the most optimal for our uses.

We hypothesize and expect to see trends that demonstrate that lower spending per pupil directly correlates to a higher incarceration rate per state. Additionally, we also expect to see the trend that a higher teacher

to pupil ratio directly correlates to a higher incarceration per state. Our research addresses the question of how education spending per pupil affects incarceration rates? and how does the teacher to pupil ratio affect the incarceration rate?

Tidying

```
#We begin by importing our data frames
prison_pop <- read.csv("ImprisonmentDataModified.csv")
teacher_pupil_ratio <- read.csv("TeacherToPupilRatioModified.csv")
dolperpupil <- read.csv("ModifiedPublicElementarySecondaryEducationFinances.csv")
total_pop <- read.csv("apportionment-2020-table02.csv")
```

```
#We begin by updating the teacher pupil ratio tibble so that the location variable only lists States, m
teacher_pupil_ratio <- teacher_pupil_ratio %>%
  rename("State" = "State.or.jurisdiction")
```

```
#We use this command to remove the District of Columbia because while it does have the teacher_pupil_ra
teacher_pupil_ratio <-teacher_pupil_ratio %>%
  filter(!row_number() %in% c(10))
```

```
#This chunk of code is done to create a function in order to automatically abbreviate the state names i
convert_to_abbreviation <- function(state_name) {
  state_index <- match(state_name, state.name)
  if (!is.na(state_index)) {
    return(state.abb[state_index])
  } else {
    return(state_name)
  }
}
```

```
#Now we apply our function to the teacher_pupil_ratio dataset in order to abbreviate all states.
teacher_pupil_ratio$State <- unlist(lapply(teacher_pupil_ratio$State, convert_to_abbreviation))
```

Now that the teacher_pupil_ratio dataset is tidied we will move on to tidying the prison_pop dataset.

In this particular dataset we also need to convert the state names to abbreviations however when we try this with the unmodified dataset we realize some of the state variables have extra characters such as "/g" at the end of the state name which the abbreviation function we created cannot handle. So first we must fix all of the states with weird extra characters at the end.

```
prison_pop$State[prison_pop$State == "Federal"] <- "United States"
```

```
prison_pop$State[prison_pop$State == "Idaho/d"] <- "Idaho"
```

```
prison_pop$State[prison_pop$State == "Iowa/e"] <- "Iowa"
```

```
prison_pop$State[prison_pop$State == "Kentucky/e"] <- "Kentucky"
```

```
prison_pop$State[prison_pop$State == "Maryland/f"] <- "Maryland"
```

```
prison_pop$State[prison_pop$State == "Massachusetts/g"] <- "Massachusetts"

prison_pop$State[prison_pop$State == "Montana/e"] <- "Montana"

prison_pop$State[prison_pop$State == "New Mexico/h"] <- "New Mexico"

prison_pop$State[prison_pop$State == "Pennsylvania/f"] <- "Pennsylvania"

prison_pop$State[prison_pop$State == "Rhode Island/d,f"] <- "Rhode Island"

prison_pop$State[prison_pop$State == "South Carolina/i"] <- "South Carolina"

prison_pop$State[prison_pop$State == "South Dakota/i"] <- "South Dakota"

prison_pop$State[prison_pop$State == "Virginia/c"] <- "Virginia"

#This function grabs the state variable from the prison_pop tibble, and changes it so that the names ar
prison_pop$State <- state.abb[match(prison_pop$State, state.name)]
```

Now we will tidy the 'dolperpupil' data set.

```
#We begin by modifying the tibble in order to rename the columns at the top (X.1,X.2, etc.) into their
dolperpupil <- dolperpupil %>%
  rename("State" = "Elementary.secondary.revenue",
         "total_revenue" = "X",
         "federal_revenue" = "X.2",
         "state_revenue" = "X.4",
         "local_revenue" = "X.6",
         "total_spending" = "X.7",
         "total_instruction_spending" = "X.9",
         "total_salary_spending" = "X.11",
         "total_employee_benefits" = "X.13",
         "total_general_administration" = "X.15",
         "total_school_administration" = "X.17")

#Now we will be subsetting the data to remove uncessary columns
dolperpupil = subset(dolperpupil, select = -c(X.1, X.3, X.5, Current.spending, X.8, X.10, X.12, X.14, X

#we use this command to remove the top three rows as they do not contain any relavent data for our rese
dolperpupil <- dolperpupil %>%  filter(!row_number() %in% c(1, 2, 3))

#Some state variable contained the text "/t" after the state abbreviation which interferes with our abi
dolperpupil$State <- gsub("\t", "", dolperpupil$State)
```

For this Section we had to rename the state variables to the full state name so they can all use the function we declared earlier to rename states as their abbreviation. **WARNING THIS IS TEDIOUS**

```
dolperpupil$State[dolperpupil$State == "US........."] <- "United States"

dolperpupil$State[dolperpupil$State == "NY........."] <- "New York"

dolperpupil$State[dolperpupil$State == "TX........."] <- "Texas"

dolperpupil$State[dolperpupil$State == "UT........."] <- "Utah"

dolperpupil$State[dolperpupil$State == "CA........."] <- "California"

dolperpupil$State[dolperpupil$State == "DC........."] <- "District of Columbia"

dolperpupil$State[dolperpupil$State == "DC........."] <- "District of Columbia"

dolperpupil$State[dolperpupil$State == "VT........."] <- "Vermont"

dolperpupil$State[dolperpupil$State == "UT........."] <- "Utah"

dolperpupil$State[dolperpupil$State == "CT........."] <- "Connecticut"

dolperpupil$State[dolperpupil$State == "NJ........."] <- "New Jersey"

dolperpupil$State[dolperpupil$State == "MA........."] <- "Massachusetts"

dolperpupil$State[dolperpupil$State == "PA........."] <- "Pennsylvania"

dolperpupil$State[dolperpupil$State == "NH........."] <- "New Hampshire"

dolperpupil$State[dolperpupil$State == "IL........."] <- "Illinois"

dolperpupil$State[dolperpupil$State == "RI........."] <- "Rhode Island"

dolperpupil$State[dolperpupil$State == "WY........."] <- "Wyoming"

dolperpupil$State[dolperpupil$State == "ME........."] <- "Maine"

dolperpupil$State[dolperpupil$State == "DE........."] <- "Delaware"

dolperpupil$State[dolperpupil$State == "AK........."] <- "Alaska"

dolperpupil$State[dolperpupil$State == "MD........."] <- "Maryland"

dolperpupil$State[dolperpupil$State == "WA........."] <- "Washington"
```

```
dolperpupil$State[dolperpupil$State == "HI........."] <- "Hawaii"

dolperpupil$State[dolperpupil$State == "ND........."] <- "North Dakota"

dolperpupil$State[dolperpupil$State == "MN........."] <- "Minnesota"

dolperpupil$State[dolperpupil$State == "MI........."] <- "Michigan"

dolperpupil$State[dolperpupil$State == "OR........."] <- "Oregon"

dolperpupil$State[dolperpupil$State == "OH........."] <- "Ohio"

dolperpupil$State[dolperpupil$State == "WI........."] <- "Wisconsin"

dolperpupil$State[dolperpupil$State == "IA........."] <- "Iowa"

dolperpupil$State[dolperpupil$State == "LA........."] <- "Louisiana"

dolperpupil$State[dolperpupil$State == "SC........."] <- "South Carolina"

dolperpupil$State[dolperpupil$State == "MT........."] <- "Montana"

dolperpupil$State[dolperpupil$State == "KS........."] <- "Kansas"

dolperpupil$State[dolperpupil$State == "VA........."] <- "Virginia"

dolperpupil$State[dolperpupil$State == "NE........."] <- "Nebraska"

dolperpupil$State[dolperpupil$State == "WV........."] <- "West Virginia"

dolperpupil$State[dolperpupil$State == "CO........."] <- "Colorado"

dolperpupil$State[dolperpupil$State == "MO........."] <- "Missouri"

dolperpupil$State[dolperpupil$State == "NM........."] <- "New Mexico"

dolperpupil$State[dolperpupil$State == "GA........."] <- "Georgia"

dolperpupil$State[dolperpupil$State == "IN........."] <- "Indiana"

dolperpupil$State[dolperpupil$State == "KY........."] <- "Kentucky"

dolperpupil$State[dolperpupil$State == "SD........."] <- "South Dakota"
```

```
dolperpupil$State[dolperpupil$State == "AR........."] <- "Arkansas"

dolperpupil$State[dolperpupil$State == "AL........."] <- "Alabama"

dolperpupil$State[dolperpupil$State == "FL........."] <- "Florida"

dolperpupil$State[dolperpupil$State == "NV........."] <- "Nevada"

dolperpupil$State[dolperpupil$State == "TN........."] <- "Tennessee"

dolperpupil$State[dolperpupil$State == "MS........."] <- "Mississippi"

dolperpupil$State[dolperpupil$State == "OK........."] <- "Oklahoma"

dolperpupil$State[dolperpupil$State == "AZ........."] <- "Arizona"

dolperpupil$State[dolperpupil$State == "NC........."] <- "North Carolina"

dolperpupil$State[dolperpupil$State == "ID........."] <- "Idaho"

#Because we do not have the appropriate data, the District of Columbia will not be included so we remov
dolperpupil <- dolperpupil %>%  filter(!row_number() %in% c(2))

#Finally we take all the state names and convert them to their abbreviations in order to match the othe
dolperpupil$State <- unlist(lapply(dolperpupil$State, convert_to_abbreviation))

#Now within the total populatin dataset we need to change the area variable name to state in order to m
total_pop <- total_pop %>%
  rename("State" = "AREA") %>%
  filter(!row_number() %in% c(9))

#Now we need to tidy the U.S. Population dataset
total_pop$State <- unlist(lapply(total_pop$State, convert_to_abbreviation))
```

Joining/Merging Now that all of our datasets are tidy it is time to merge them into one data set with all variables.

Before joining, all data sets had 51 observations which consisted on a row for each state in the United States and a row for the United States. The common ID among all data sets was the state column which consisted of each individual state name and a observation for the United States. Therefore, joining all data sets by the common ID state made since, ultimately, leading to no row being dropped after joining the data sets.

```
#We will begin by naming a new dataset, just called dataset, then joining the prison_pop tibble to teac
dataset <- teacher_pupil_ratio %>%
  left_join(prison_pop, by = "State")

#creating an updated dataset which includes the dolperpupil set joined by their shared state variable
dataset <- dataset %>%
  left_join(dolperpupil, by = "State")
```

```r
#Like the work above, just updating the dataset to include total_pop by their shared state variable
dataset <- dataset %>%
  left_join(total_pop, by = "State")

#renaming the total column to total prisoners makes the dataset easier to understand
dataset <- dataset %>%
  rename("total_prisoners" = "Total")

#This command converts the total prisoner variable into a numeric variable
dataset$total_prisoners <- as.numeric(gsub(",", "", dataset$total_prisoners))

#This command converts the total revenue variable into a numeric variable
dataset$total_revenue <- as.numeric(gsub(",", "", dataset$total_revenue))

#Now we are going to abbreviate the United States as US so it is abbreciated like the state names
dataset$State[dataset$State == "United States"] <- "US"

#Now we are going to remove the US from our dataset as it is not
dataset <- dataset[dataset$State != "US", ]
```

Wrangling

```r
#This command converts the totalspending variable into a numeric variable
dataset$total_spending <- as.numeric(gsub(",", "", dataset$total_spending))

#Now we want to rename the population variable to make it easier to work with
dataset <- dataset %>%
  rename("total_population" = "RESIDENT.POPULATION..APRIL.1..2020.")

#This command will convert total populaion from a character to a numeric variable
dataset$total_population <- as.numeric(gsub(",","", dataset$total_population))

# the mutate function creates a new categorical varibale "spending_category" that assigns value consist
dataset <- dataset %>%
  mutate(spending_category = cut(total_spending,
                                 breaks=c(0, 8333, 16000, 25000),
                                 labels=c("Low", "Medium", "High")))

#This code selects just the relevant variables then arranges them by descingnd amount of per pupil spen
data2 <- dataset %>%
dplyr::select(State, Pupil.Teacher.Ratio,total_spending, total_prisoners, total_population) %>%
  arrange(desc(total_spending))

#In this code chunk we are using the mutate command to create a new variable for incarceration rate as
data2 <- data2 %>%
  mutate(incarceration_rate = total_prisoners/total_population * 100)

#Now using the summarise command we can obtain the average spending, incarceration rate and pupil to te
data2 %>%
  summarise("Average pupil-teacher ratio" = mean(Pupil.Teacher.Ratio),"Average total spending" = mean(t
  )
```

```
##   Average pupil-teacher ratio Average total spending Average incarceration rate
## 1                      15.466               14438.12                   0.3275589
```

From our Summary Statistics we have found the average pupil to teacher ratio across the U.S. is 15.46 Students per teacher. On average schools spend $14,438.12 per pupil. The average incarceration rate across the U.S. is 0.32% of the population.

```
#This command is to rename the state function to lowercase so the following plots will work
dataset <- dataset %>%
  rename("state" = "State")
```

```
#This command is to rename the state function to lowercase so the following plots will work
data2 <- data2 %>%
  rename("state" = "State")
```

```
#In this code chunk we want to use the mutate command to create a new categorical variable for the regi
west = c("WA", "OR", "CA", "NV", "AZ", "ID", "MT", "WY",
          "CO", "NM", "UT")
  south = c("TX", "OK", "AR", "LA", "MS", "AL", "TN", "KY",
           "GA", "FL", "SC", "NC", "VA", "WV")
  midwest = c("KS", "NE", "SD", "ND", "MN", "MO", "IA", "IL",
             "IN", "MI", "WI", "OH")
  northeast = c("ME", "NH", "NY", "MA", "RI", "VT", "PA",
              "NJ", "CT", "DE", "MD", "DC")
data2 <- data2 %>%
  mutate(
    region = ifelse(state %in% west, "west", NA),
   region = ifelse(is.na(region) & state %in% northeast, "northeast", region),
   region = ifelse(is.na(region) & state %in% midwest, "midwest", region),
   region = ifelse(is.na(region) & state %in% south, "south", region),
   region = ifelse(is.na(region), "other", region)
  )
```

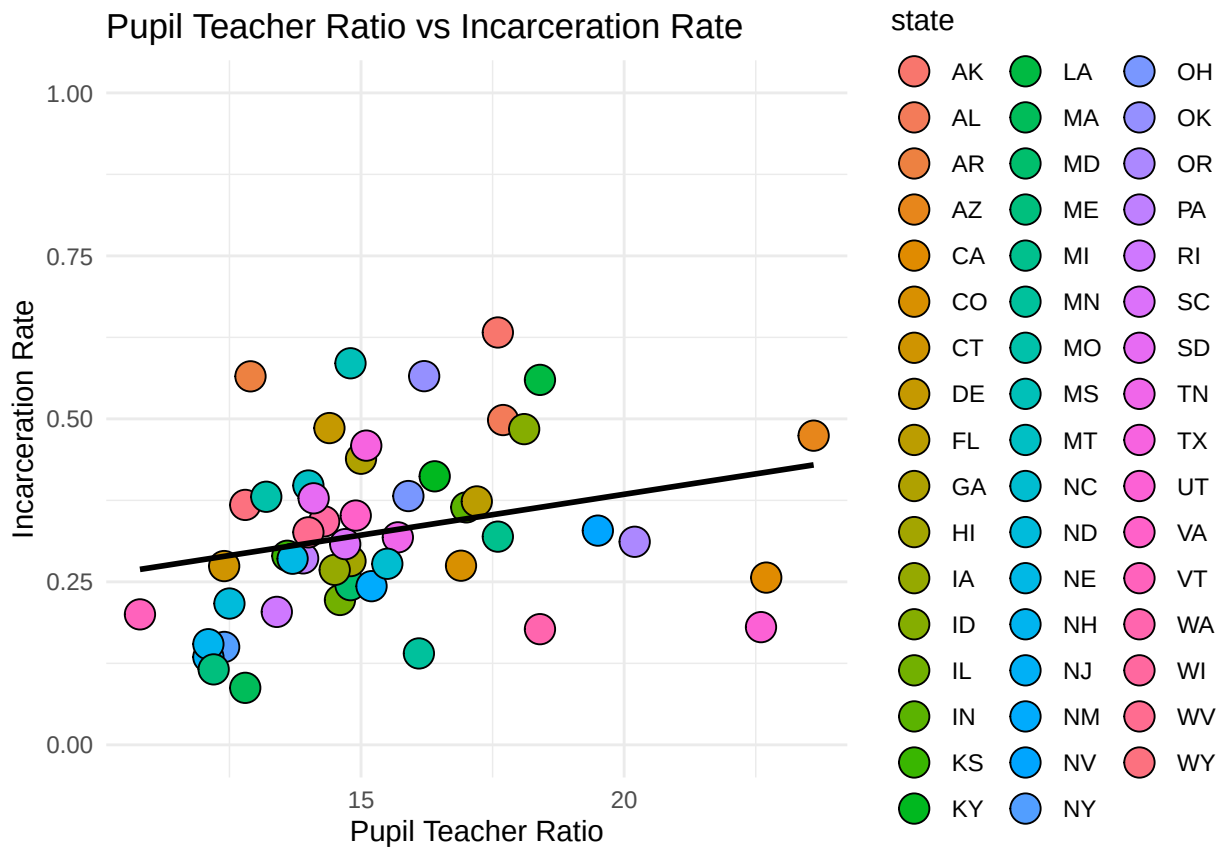
```
data2 %>%
  group_by(region) %>%
  summarise("Average pupil-teacher ratio" = mean(Pupil.Teacher.Ratio),"Average total spending" = mean(t
```

```
## # A tibble: 5 x 4
##   region    Average pupil-teache~1 Average total spendi~2 Average incarceratio~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 midwest                     14.8                 13741.                  0.299
## 2 northeast                   12.8                 20257.                  0.213
## 3 other                       16.2                 16294                   0.457
## 4 south                       15.6                 11484.                  0.431
## 5 west                        18.5                 12802.                  0.318
## # i abbreviated names: 1: `Average pupil-teacher ratio`,
## #   2: `Average total spending`, 3: `Average incarceration rate`
```

Visualizing

```r
#The ggplot() + geom_point() function creates a scatter plot of the distribution between pupil teacher
data2 %>%
  ggplot(aes(x = Pupil.Teacher.Ratio, y = incarceration_rate)) +
  geom_point(aes(fill = state), shape = 21, size = 5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  theme_minimal() +
  labs(title = "Pupil Teacher Ratio vs Incarceration Rate",
       x = "Pupil Teacher Ratio",
       y = "Incarceration Rate") +
  theme(legend.position = "right") +
  scale_y_continuous(limits = c(0, 1))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

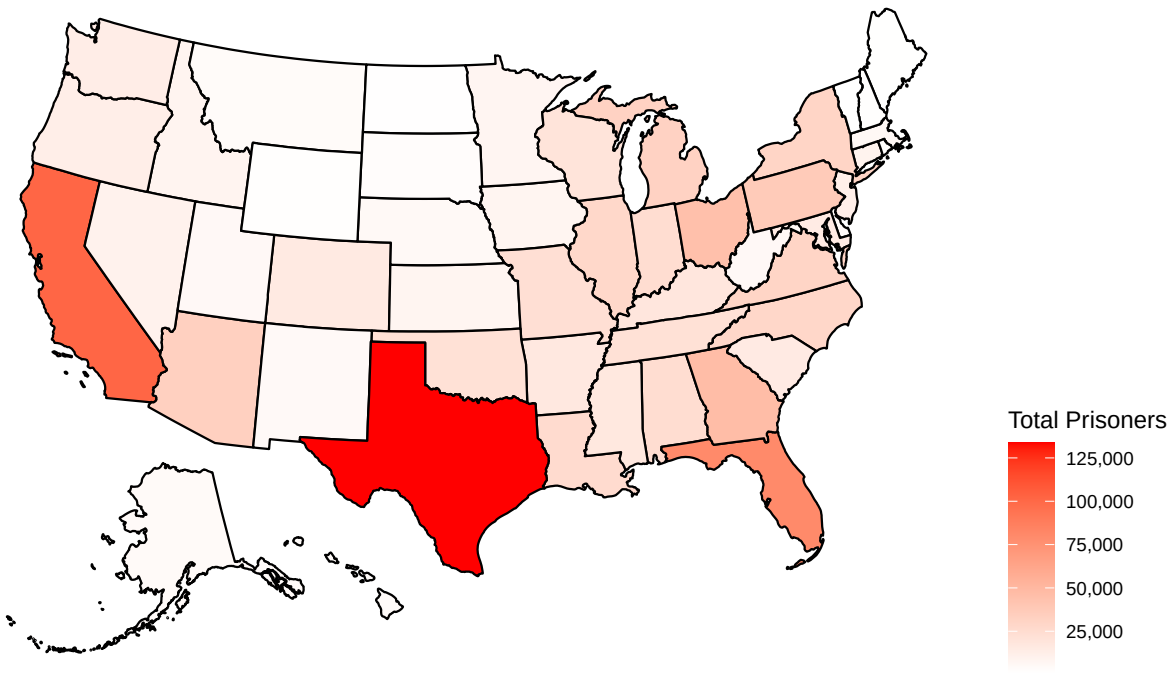


Graph 1: Giovanni This scatter plot displays the "Pupil Teacher Ratio" along the x-axis and the "Incarceration Rate" along the y-axis. Ultimately, this scatter plot demonstrates the relationship between the pupil teacher ratio and the incarceration rate and the general trend indicates that as the pupil teacher ratio increases so does the incarceration rate across the US states.

```r
#The plot_usmap() function creates a map of the US states that demonstrates differing values of prisone

plot_usmap(data = dataset, values = "total_prisoners") +
  scale_fill_continuous(name = "Total Prisoners",
                        label = scales::comma,
                        low = "white",
```

```
                        high = "red") +
  theme(legend.position = "right")
```



Graph 2: Giovanni This US map plot demonstrates the total population of prisoners across each US state. With a lighter red indicating a low prisoner population and a vibrant red indicating a high prisoner count.

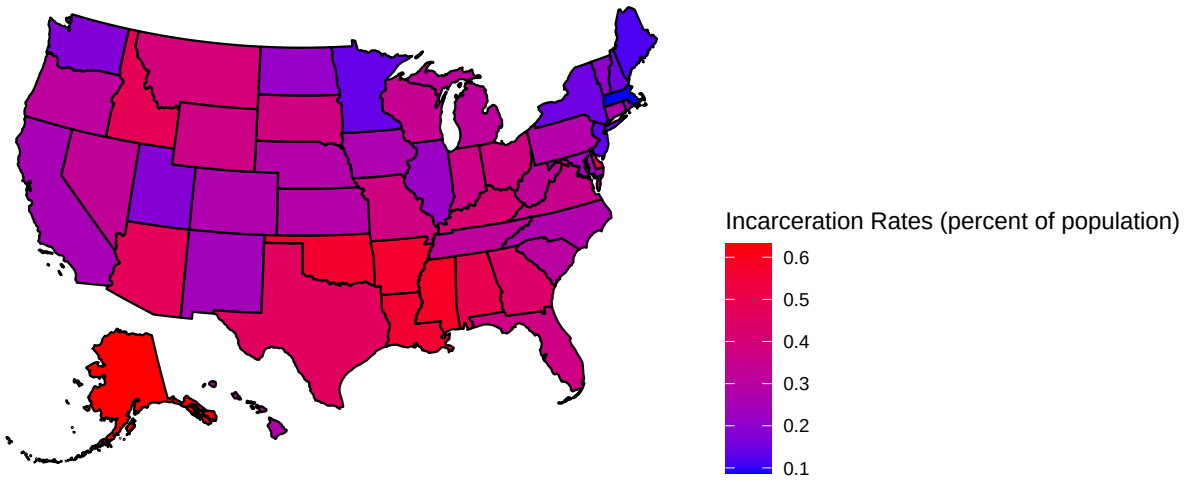
```
# the pivot_longer() function assigns a total_prisoners() value to correspond with the newly created sp
dataset %>%
  pivot_longer(cols = c(total_prisoners),
               values_to = "Value") %>%
  ggplot(aes(x = spending_category, y = Value, fill = spending_category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total Prisoners vs Spending Category", y = "Total Prisoners", x = "Spending Category\n\n
  scale_y_continuous(breaks = seq(0, 140000, 10000), limits = c(0, 140000)) +
  theme_minimal()
```

(Low: 0−$8333, Medium: $8333−$16000, High: $16000−$25000)

Graph 3: Giovanni This bar plot demonstrates the total number of prisoners based upon the spending per pupil categories of (Low, Medium, and High). This graph demonstrates that states with a "Medium" spending per pupil contribute to approximately 35,000 more prisoners than states with a "High" spending per pupil.

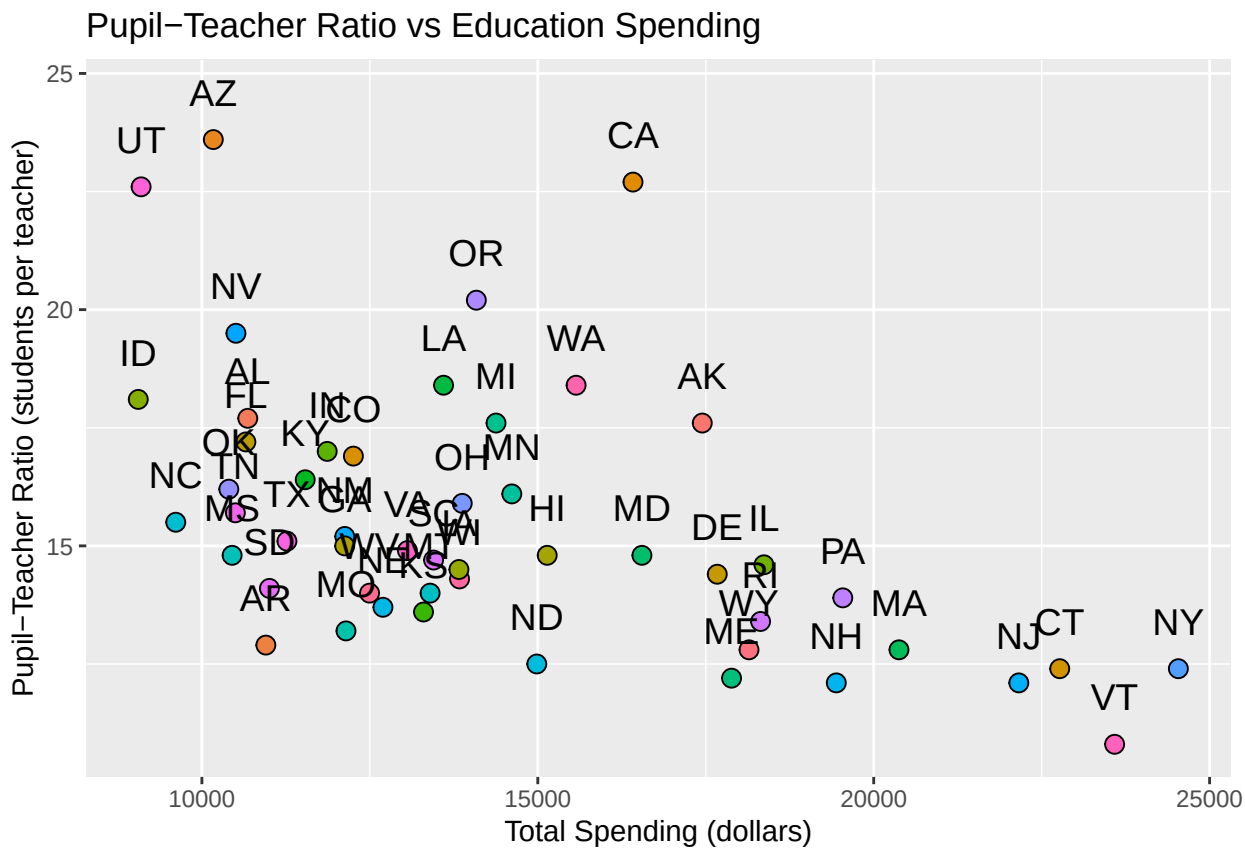
```
#Using the us map plot for data2, set the values to equal the states incarceration rates, use scale fil

plot_usmap(data = data2, values = "incarceration_rate") +
  scale_fill_continuous(name = "Incarceration Rates (percent of population)", low = "blue", high = "red
  theme(legend.position = "right")
```



Graph 1: Michael This graph shows the incarceration rates as a percentage of the population for all 50 states. THis graph shows increased incarceration in the South-East of the U.S. and lower incarceration rates in the North-East.

```
#port the data2, use ggplot to make my plot with the x axis being total spending and pupil-teacher rati
data2 %>%
  ggplot(aes(x = total_spending, y = Pupil.Teacher.Ratio, fill = state, label = state)) +
  geom_point(size = 3, shape = 21) +
  geom_text(size = 5, nudge_y = 1) +
  labs(title = "Pupil-Teacher Ratio vs Education Spending", x = "Total Spending (dollars)", y = "Pupil-
```



Graph 2: Michael This graph shows a comparison each states total spending per pupil in comparison to its pupil to teacher ratio showing some correlation between greater spending per pupil and smaller class sizes.

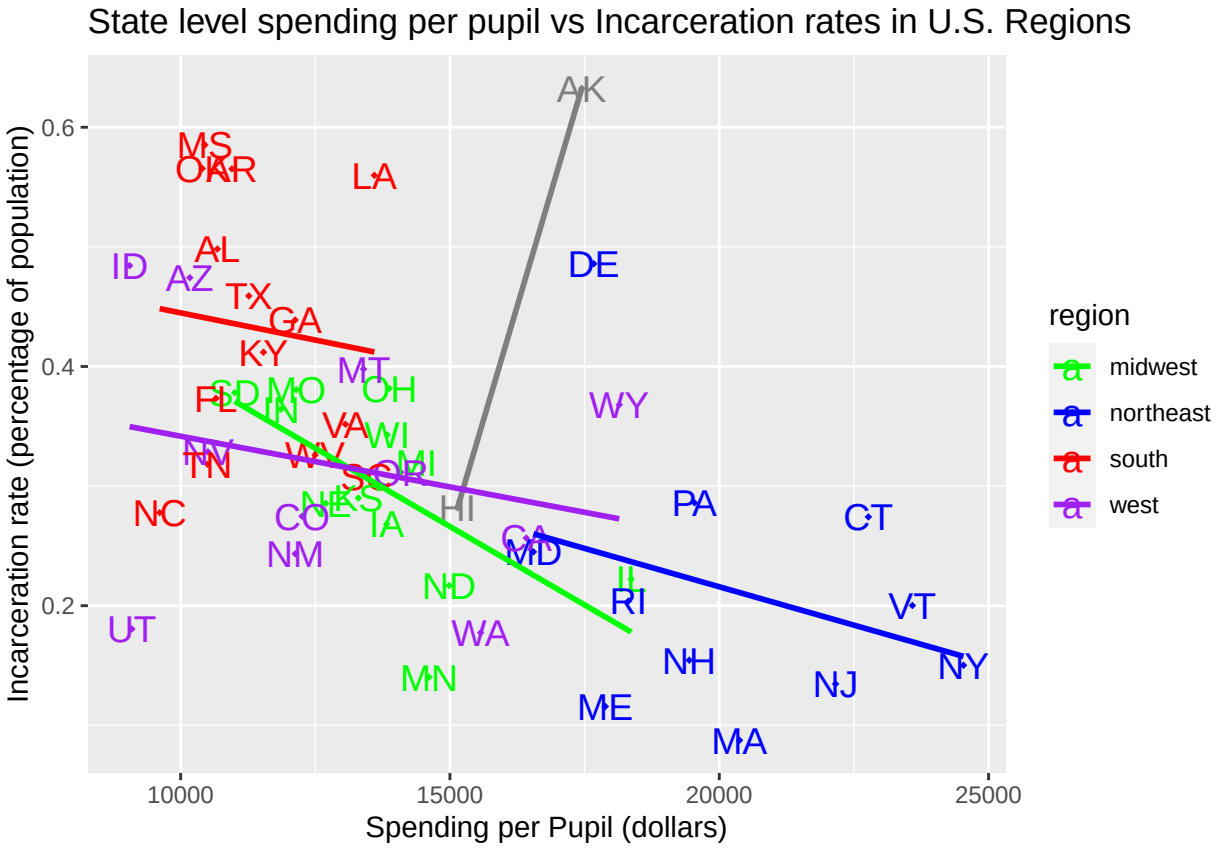
```
#port the data, use ggplot to start, set the variables, use the label = state function to assign each p

region_colors <- c("northeast" = "blue", "midwest" = "green", "south" = "red", "west" = "purple")

data2 %>%
  ggplot(aes(x = total_spending, y = incarceration_rate, color = region, label = state)) +
  geom_point(shape = 18, size = 1) +
  geom_text(size = 5, nudge_x = 0.5) +
  geom_smooth(aes(group = region), method = "lm", se = FALSE) +
  scale_color_manual(values = region_colors) +
  labs(title = "State level spending per pupil vs Incarceration rates in U.S. Regions", x = "Spending p
  theme(legend.position = "right")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: label
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Graph 3: Michael These five graphs show a the relationship between states spending per pupil and their incarceration rate. These graphs are sorted by region in order to see if the relationship is similar in all areas of the country which seems to hold. My prediction that states with greater spending per pupil see lower in carceration however the effect is more dramatic in the midwest when compared to the south. It should also be mentioned the relationship between Alaska and Hawaii should not be considered as it is only two data points.

Discussion

Michael: Research Question My hypothesized relationship between states that spend more per pupil in general see lower incarceration rates did in fact hold true. If you look at "Graph3: Michael" you will see that although the slope of the line of best fit changed, each region had the same negatively sloped line.

Michael: Reflection This process was difficult yet when my code started working and graphs were actually being produced I felt elated. One thing I did discover is how difficult it is to code as a team, because each minor detail in each of our codes made a world of difference. For example I named my state variable "state", Giovanni labeled his "State" which meant I spent quite some time combing through our document for consistency in case.

Giovanni: Research Question I hypothesized that the relationship between teacher to pupil ratio and incarceration rate will indicate that a higher teacher to pupil ratio will correlate with a higher incarceration rate and vice versa. "Graph 1: Giovanni" demonstrates a moderate positive linear relationship as demonstrated by the linear regression line, an increase in the teacher to pupil ratio leads to an increase in the incarceration rate.

Giovanni: Reflection The process of finding the adequate graph to represent your data and demonstrate your desired results is thought invoking and time consuming. This is because the process required extensive data modification and experimentation through trial and error to create a working visualization.