Finding the Relationship between Federal Revenue for Public Schools and Violent Crimes, as well as Median Household Income and Violent Crimes

Aileen Li, Nyah Strickland

2023-03-04

Many U.S. states have made changes to public school funding. One of our research questions is if there is a relationship between change in federal revenue for public schools and violent crimes by state during 1992-2016. We wanted to explore this because if there is a relationship, depending on what kind, states can make changes in their budget for the betterment of their enrolled students. We expected the more funding for public schools, the less violent crimes there are by state.

As inflation increases annually, household incomes must also rise to adjust for standard of living in each state. Our second research question is if there is a relationship between median household income and violent crimes by state during 1992-2016. We wanted to explore this because if there is a relationship between these two datasets, then state governments should make changes in their policies to help states where households are most affected. We expect the greater the median household income, the less violent crimes there are by state. The following are the research questions we investigated:

-Is there a relationship between Federal Revenue for Public Schools and Crimes?
-Is there a relationship between Median Household Income and Crimes?

And the following are the datasets we used(click to access):
-Median Household Income by State
-U.S. Education Dataset
-Crime Rate by State Dataset

The median household income, U.S. Education, and state crime rate datasets were acquired from the U.S. Census Bureau, U.S. Census Bureau and the National Center for Education Statistics (NCES), and Unified Crime Reporting Statistics and under the collaboration of the U.S. Department of Justice and the Federal Bureau of Investigation, respectively. We joined the data sets by the state names, which are the unique rows, and year. Specifically, we looked into the years from 1992 to 2016 for both research questions. For the median household income dataset, the unique variable is the median household income (Median income) in the current dollars dataset, measuring the annual state median household incomes without considering inflation. As for the U.S. Education dataset, the unique variable is federal revenue (FEDERAL_REVENUE), and for the violent crime rate dataset, the unique variables are violent crime rate (Data.Rates.Violent.All), nonviolent crime rate (Data.Rates.Property.All), burglary crime rate (Data.Rates.Property.Burglary), and murder rate (Data.Rates.Violent.M The key variables are categorical while all the unique variables are numerical. These are the eight variables required for the project.

First, we loaded the libraries and oeiginal datasets we needed.

```
library(tidyverse)
library(readr)
library(readxl)
states_all_extended <- read_csv("states_all_extended.csv")
state_crime <- read_csv("state_crime.csv")
h08_1_ <- read_excel("h08 (1).xls")
```

Tidying

We then started tidying our datasets. For all the datasets, it is important to note we did not include Washington D.C as a state.

DESCRIBE

```
# Only keep key and unique variables for each dataset
# Rename dataset to us_education
us_education <- states_all_extended %>%
  select('STATE', 'YEAR', 'FEDERAL_REVENUE') %>%
  filter(!str_detect(STATE, "^DISTRICT")) %>%
  rename("State" = "STATE", "Year" = "YEAR",  "Federal_Revenue" = "FEDERAL_REVENUE") %>%
# Make sure state names are identical in both datasets
  mutate(State = str_replace(State, "_", " "))
```

DESCRIBE

```
# Rename crime variable names so they don't have the word 'Data' in it
crime_rate <- state_crime %>%
  select('State',
       'Year',
       'Data.Rates.Violent.All',
       'Data.Rates.Property.All',
       'Data.Rates.Property.Burglary',
       'Data.Rates.Violent.Murder') %>% filter(!str_detect(State, "^District")) %>%
  rename("All_Property_Rates" = "Data.Rates.Property.All",
         "All_Property_Burglary_Rates" = "Data.Rates.Property.Burglary",
         "All_Violent_Rates" = "Data.Rates.Violent.All",
         "All_Murder_Rates" = "Data.Rates.Violent.Murder") %>%
# Put year in ascending order so it is easy to merge this dataset with us_education
  arrange(Year)
# Make state names be uppercase to make merging the datasets easier
crime_rate$State <- toupper(crime_rate$State)
```

The median_housing_income dataset is untidy and contains unnecessary info in relation to what we wanted to analyze, so we modified it for our questions.
We changed the column names in the dataset to be named in sequential order. Then, we deleted all the columns containing standard error data, which was every other column starting at the second column.

```
# Find standard error columns and delete them
med_income_base = h08_1_
colnames(med_income_base) = paste(seq_along(med_income_base))
del_col = list()
for (col in 1:ncol(h08_1_)) {
  if (col!=1 & (col-1)%%2==0) {
    del_col= append(del_col, as.character(col))
  }
}
med_income_base = med_income_base[,!names(med_income_base) %in% del_col]
```

We renamed the columns to their respective titles given in the dataset on the fourth row. And then deleted rows containing unimportant information. There were two columns containing different info for the year 2013. We tried to find why they were different, in addition as to what the values in the parentheses meant, but there was no documentation for this dataset. Thus, we decided to keep the first column instance of the year 2013 so that we would have no differing data for the same year and state combination that could possibly affect our graphs.

```
# Rename columns
colnames(med_income_base) = med_income_base[4,]

# Remove undesired data
med_income_base = med_income_base[-(1:6),-9]
i = min(which(med_income_base$State=='Wyoming'))
med_income_untidy = med_income_base[1:i,] |> filter(State!="D.C.")
```

We made the median_housing_income dataset tidy by changing the individual year columns to be under the Year variable and their values under Median_Income. Then, we filtered for observations within our desired time period, which is from 1992 to 2016, and arranged them in ascending order.

```
# Make desired data tidy
med_income_tidy = med_income_untidy |>
  pivot_longer(!State, names_to="Year", values_to="Median_Income") |>
  filter(substr(Year, 1, 4) %in% (1992:2016)) |> arrange(Year)
med_income_tidy$State = toupper(med_income_tidy$State)
```

We found years with values in parentheses and checked for duplicate years for the same state. We also removed the values in parentheses for the rest of the years. And for the sake of merging the datasets easier, we completely capitalized the state names.

```
# Find years with () and duplicate years other than 2013
pyears = med_income_tidy[grep(")$",med_income_tidy$Year),]
dupl_years = pyears |>
  mutate(compare=substr(Year,1,6)) |>
  select(State, compare) |> group_by(State, compare) |>
  filter(duplicated(compare))

# Remove () from some years
med_income_tidy$Year = substr(med_income_tidy$Year, 1,4)
med_income = med_income_tidy
```

Joining/Merging

We found the number of rows and columns in each other datasets before merging.

```
# Note how many observations are in each data set
nrow(us_education)
```

```
## [1] 1682
```

```
ncol(us_education)
```

```
## [1] 3
```

```
nrow(crime_rate)
```

```
## [1] 3055
```

```
ncol(crime_rate)
```

```
## [1] 6
```

```
nrow(med_income)
```

```
## [1] 1250
```

```
ncol(med_income)
```

```
## [1] 3
```

There are 1,682 observations and 1 unique variable (Federal_Revenue) in the us_education dataset, 3,055 observations in the state_crime dataset observations and 4 unique variables (All_Property_Rates, All_Property_Burglary_Rates, All_Violent_Rates, All_Murder_Rates), and 1,250 observations and 1 unique variable (Median_Income).

We first joined the us_education and crime_rate and then joined their merged result to med_income.

```
# Left join state_crime to us_education by key variable Year and State
rqs <- left_join(us_education,crime_rate, by = c("Year", "State")) %>%
  filter(!is.na(Federal_Revenue))

# Left join med_income to prev. joined dataset by same key var.s
med_income$Year = as.numeric(med_income$Year)
rqs = left_join(rqs, med_income, by = c("Year", "State"))
# Filter missing values of federal revenue
nrow(rqs)
```

```
## [1] 1250
```

Between us_education and state_crime, there are 2 ID variables (Year and State) in common. After joining these 2 datasets into a single dataset rqs, there are 1682 observations since the us_education spans from 1992-2016, causing the merged dataset to stop at that time frame. This is also the time period we wanted, and it is the reason we left join crime_rate to us_education rather than the other way around. After joining the med_income to the previously joined dataset, we still have 1,250 obervations but with more columns because the process added the unique columns from med_income to the observations with matching state and year in the latter, which was all of them.

There are 1,250 observations in the rqs dataset. This means that from the total number of overlapped observations (being 3055-1682=1373) in both us_education and state_crime datasets, there were 1373-1250=123 observations that had missing values. Since the crime_rate was left joined to us_education, the missing values likely arose from not matching with the year and state name for Federal_Revenue as this variable records state public school federal funding from 1992-2016.

Wrangling

We created a new variable called Fed_Rev, which is calculated by dividing the Federal_Revenue by 100,000.

```
# Create new var. from other
rqs = rqs |> mutate(Fed_Rev=Federal_Revenue/100000)
```

We then found the mean, median, maximum, and minimum for the following variables if it was possible.

```
# Compute summary stat.s
summary(rqs$State)
```

```
##    Length     Class      Mode
##      1250 character character
```

```
summary(rqs$Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1992    1998    2004    2004    2010    2016
```

There are 1,250 values of the State variable, which is a character vector. For Year in rqs, the earliest year is 1992 and most recent year is 2016. The median and average year is 2004. This makes sense sense we selected this time period and there is an equal amount of observations for each year.

```
rqs$Median_Income = as.numeric(rqs$Median_Income)
summary(rqs$Median_Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20271   36568   43642   44641   51503   76260
```

The Median_Income variable has a minimum of $20,271 and a maximum of $76,260. The mean is $44,641, and the median is $36,568.

```
rqs |> select(All_Property_Rates) |>
  summarize(mean=mean(All_Property_Rates))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1 3449.
```

```
summary(rqs$All_Property_Rates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1407    2642    3353    3449    4096    7500
```

The average rate of all reported offenses of Property-related crimes per 100,000 population by state is 3,449 people per 100,000 population. The minimum and maximum rates are 1,407 per 100,000 and 7,500 per 100,000, respectively.

```
rqs |> select(All_Property_Burglary_Rates) |>
  summarize(mean=mean(All_Property_Burglary_Rates))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  731.
```

```
summary(rqs$All_Property_Burglary_Rates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   201.7   513.9   690.0   731.4   936.5  1888.8
```

The average rate of all reported offenses of Property-related Burglaries per 100,000 population is 731. The minimum and maximum rates are 201.7 per 100,000 and 1888.8 per 100,000.

Visualizing

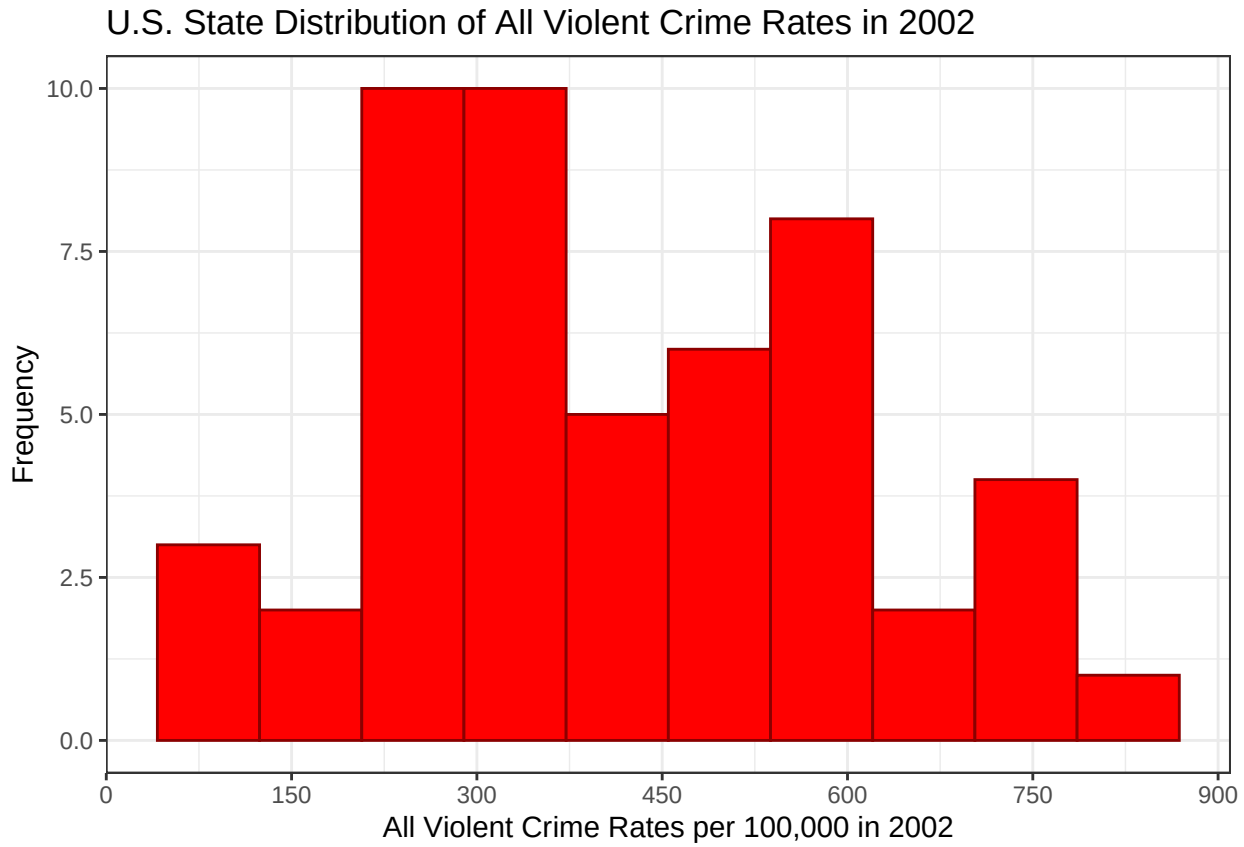
1 Variable Plots

We created plots graphing one variable each. **Figure 1** Figure 1 plots All_Violent_Rates

```
# find max violent crime rate
max(crime_rate$All_Violent_Rates)
```

```
## [1] 1244.3
```

```
# plot a histogram of all violent crime rates across the states
crime_rate |> filter(Year==2002) |>
  ggplot(aes(x= All_Violent_Rates)) +
  geom_histogram(bins=10, color ="darkred", fill = "red") +
  labs(title = "U.S. State Distribution of All Violent Crime Rates in 2002 ",y="Frequency") +
  scale_x_continuous(name = "All Violent Crime Rates per 100,000 in 2002", breaks=seq(0, 1245,150)) +
  theme_bw()
```



The U.S. state distribution of all violent crimes is right-skewed with the peak being around 300 reported

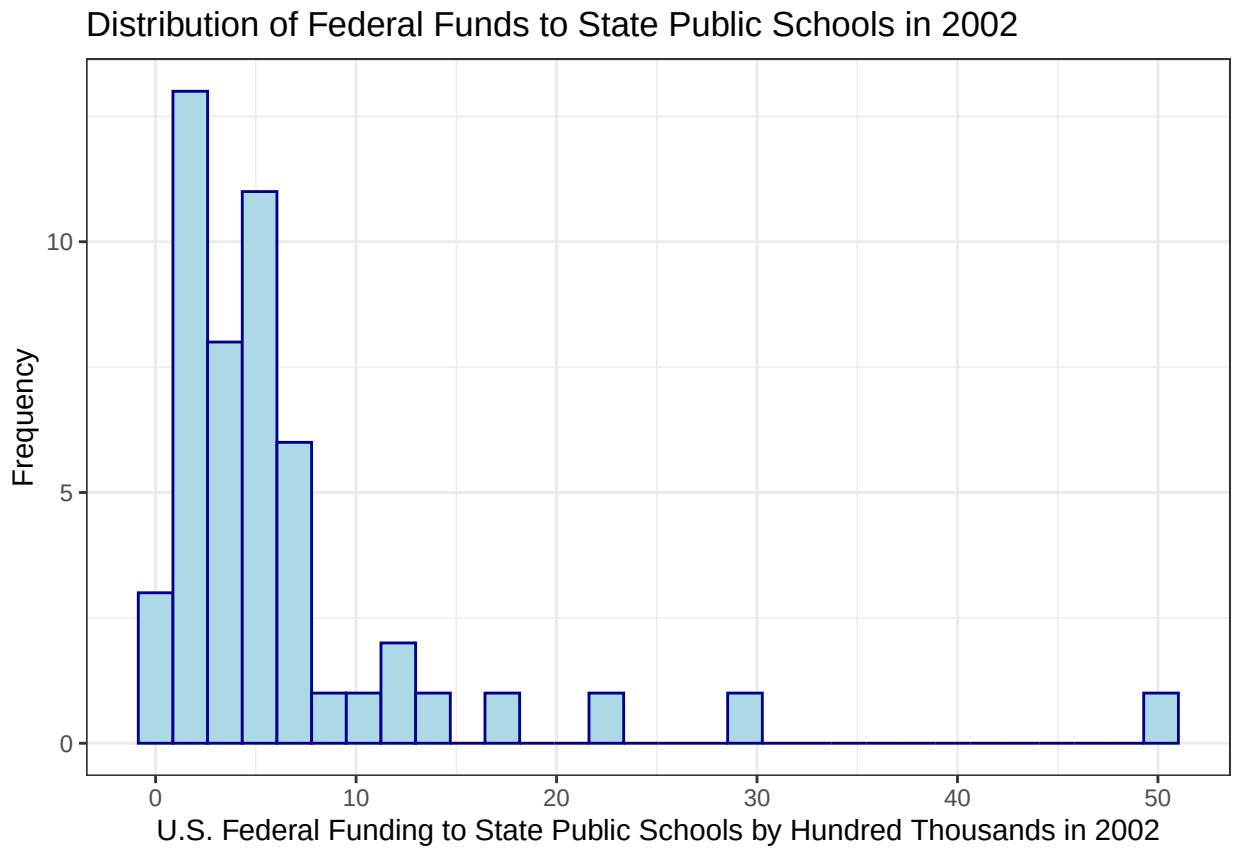
offenses per 100,000 population. Generally, violent crime rates between 150 and 450 reported offenses per 100000 population are common statewide.

Figure 2 Figure 2 plots Fed_Rev.

```
# plot U.S. federal funding to state public schools
# find max federal funds to state public schools using summary
summary(us_education$Federal_Revenue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   31020  196599  418887  780779  842409 9990221     432
```

```
# plot a histogram of U.S. federal funding to state public schools
rqs %>% filter(Year==2002) |>
  ggplot(aes(x=Fed_Rev)) +
  geom_histogram(bins=30, color ="darkblue", fill = "lightblue") +
  labs(title = "Distribution of Federal Funds to State Public Schools in 2002",
       y = "Frequency") +
  scale_x_continuous(name = "U.S. Federal Funding to State Public Schools by Hundred Thousands in 2002")
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw()
```



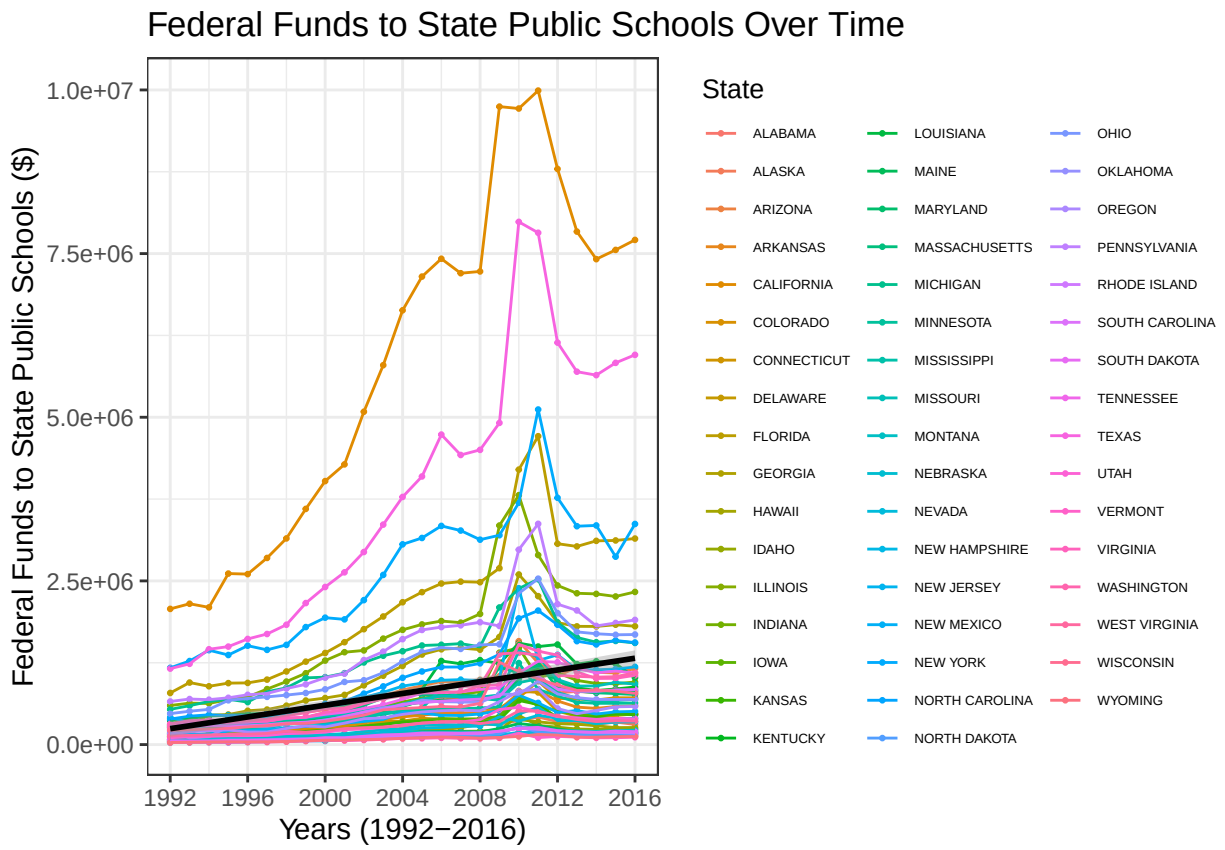
The distribution of federal funds to state public schools is strongly skewed to the right. This shows that not a lot of state public schools receive federal funds.

2 Variable Plots

We created plots graphing two variables each. **Figure 3** Figure 3 plots Year and Federal_Revenue

```
# Graph 2 variables: x = year, y = federal funds to state schools
# Create a scatterplot tracking each state's public school federal funding over time
rqs %>% ggplot(aes(x = Year, y = Federal_Revenue)) +
  geom_point(aes(color = State), size =1/2) +
  labs(title = "Federal Funds to State Public Schools Over Time",
       y = "Federal Funds to State Public Schools ($)") +
  scale_x_continuous(name = "Years (1992-2016)", breaks =seq(1992, 2016,4)) +
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw() +
# Connect each state's points with a line
  geom_line(aes(color = State)) +
  theme(legend.key.height= unit(5, 'mm'),
        legend.key.width= unit(5, 'mm'),
        legend.title = element_text(size=10),
        legend.text = element_text(size=5)) +
  # theme(legend.position = "none") +
  geom_smooth(method = "lm", color = "black")
```

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



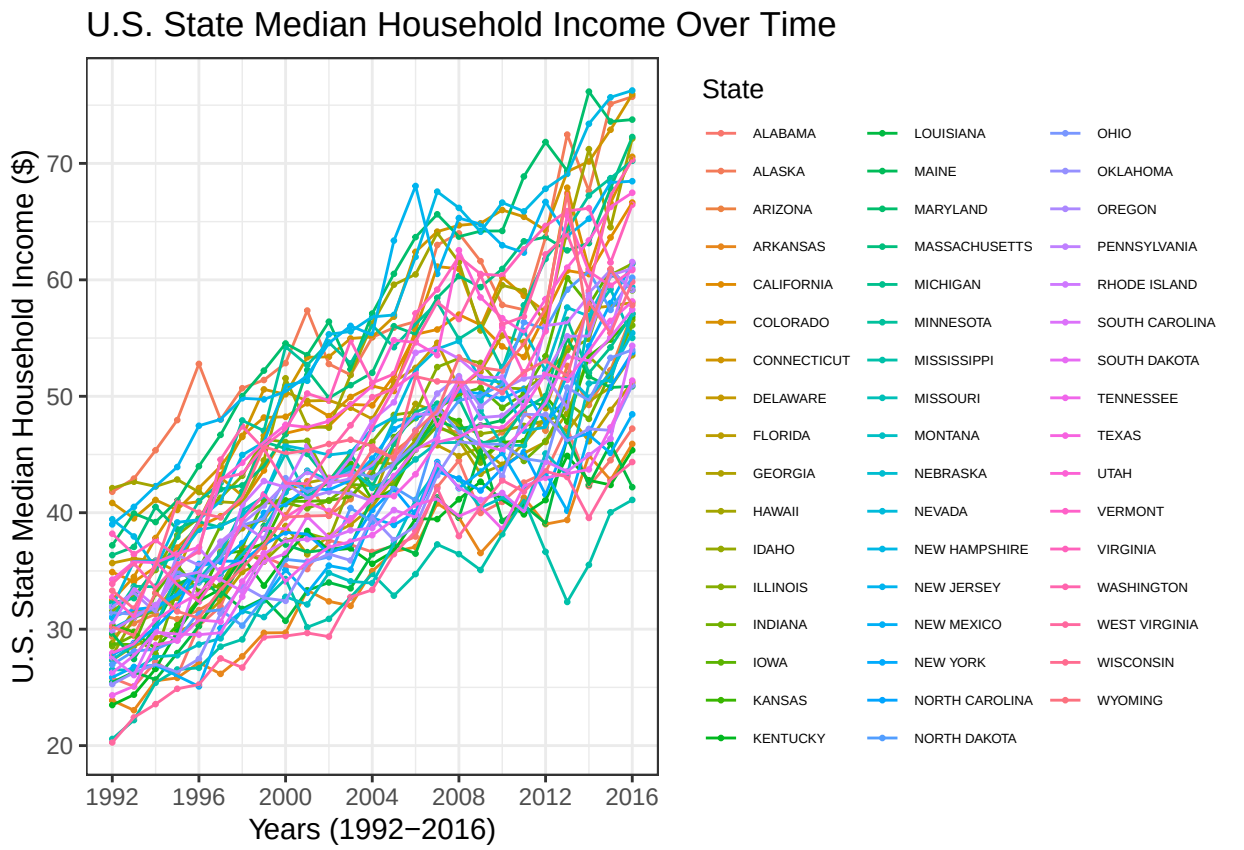
The distribution of federal funds to state public schools over 1992-2016 generally increases. Though the funds decrease between 2010 and 2012 and increase again after that time period. This decrease in funds

during 2010-2012 is likely due to the consequences of the government overspending following the U.S. 2008 recession.

Figure 4 Figure 4 plots Year and mi, which is a function of Median_Income.

```
 # Graph 2 variables: x = year, y = median household income
# Create a plot tracking each state's median household income over time
rqs = rqs |> mutate(mi=Median_Income/1000)

rqs %>% ggplot(aes(x = Year, y = mi)) +
  geom_point(aes(color = State), size =1/2) +
  labs(title = "U.S. State Median Household Income Over Time") +
  scale_y_continuous(name="U.S. State Median Household Income ($)", breaks=seq(0, 100, 10)) +
  scale_x_continuous(name = "Years (1992-2016)", breaks=seq(1992, 2016, 4)) +
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw() +
#Connect each state's points with a line
  geom_line(aes(color = State, group=State)) +
    theme(legend.key.height= unit(5, 'mm'),
        legend.key.width= unit(5, 'mm'),
        legend.title = element_text(size=10),
        legend.text = element_text(size=5))
```




```
  # theme(legend.position = "none")
```

The distribution of U.S. state median household income over 1992-2016 has a generally positive upward slope. This shows that as time goes on, state median household income also increases.

3 Variable Plots

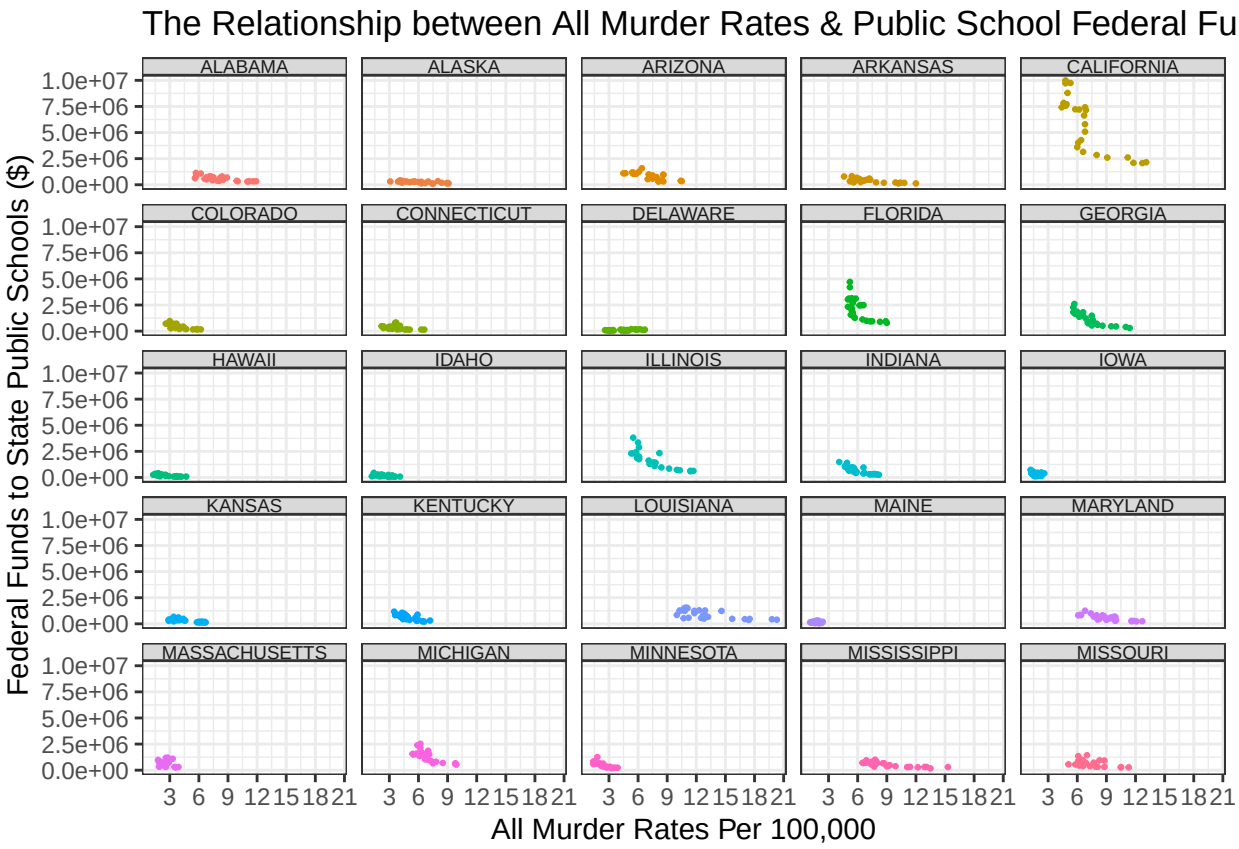
We created plots graphing three variables. We split the rsq dataset in half so that the data points on the scatterplots are more readily seen. **Figure 5** Figure 5 plots All_Murder_Rates, Federal_Revenue, and State.

```
# find max murder rates
summary(rqs$All_Murder_Rates)
```

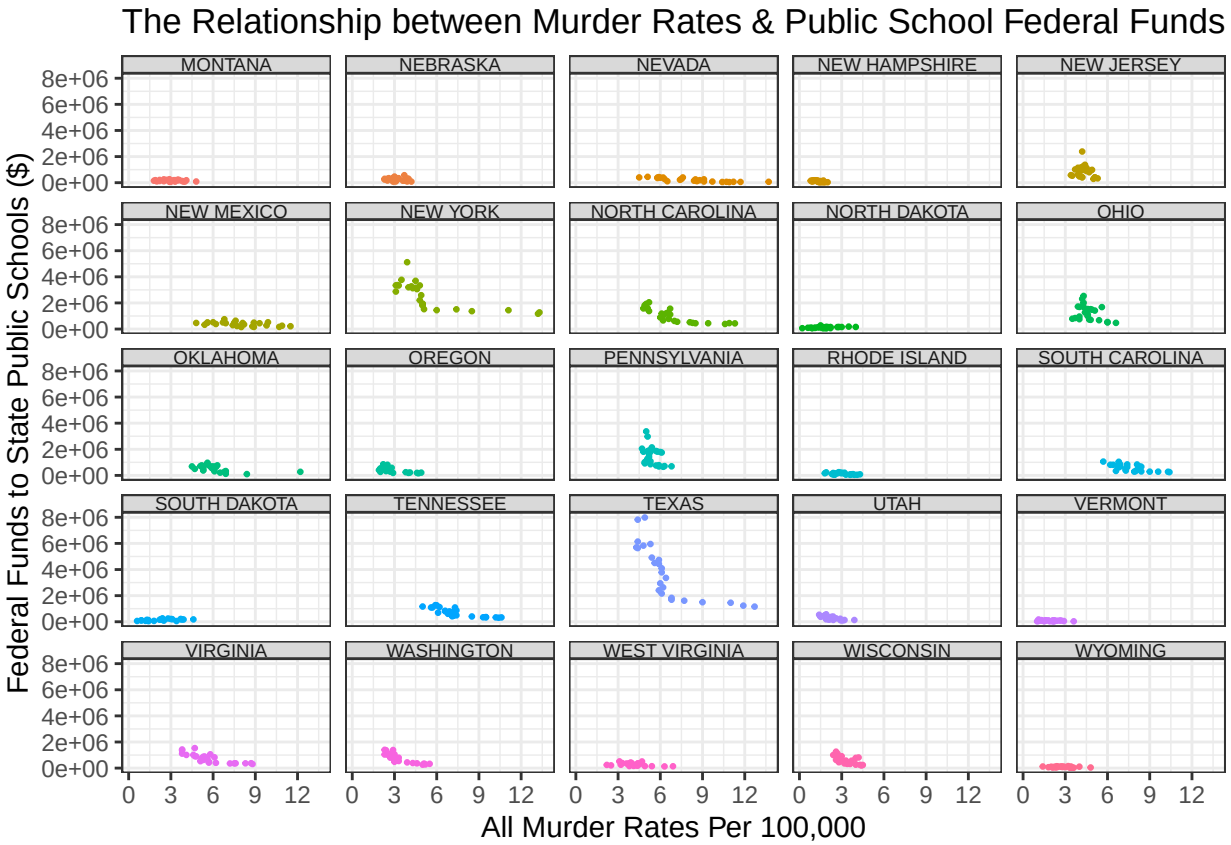
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.200   2.900   4.700   5.129   6.700  20.300
```

```
# graph 3 variables: y = federal funds to state public schools, x = crime rate (murder)
FM1 <- rqs %>% arrange(State)
FM1 <- FM1[1:625,]
FM2 <- rqs %>% arrange(State)
FM2 <- FM2[626:1250,]

# graph of first 25 states in alphabetical order
FM1 %>% ggplot(aes(x = All_Murder_Rates, y = Federal_Revenue)) +
  geom_point(aes(color = State), size =1/2) +
  labs(title = "The Relationship between All Murder Rates & Public School Federal Funds",
       y = "Federal Funds to State Public Schools ($)") +
  scale_x_continuous(name = "All Murder Rates Per 100,000", breaks =seq(0, 21,3)) +
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw() + facet_wrap(vars(State)) +
  theme(legend.position = "none", strip.text = element_text(size = 7, margin=margin()))
```




```
# graph of last 25 states in alphabetical order
FM2 %>% ggplot(aes(x = All_Murder_Rates, y = Federal_Revenue)) +
  geom_point(aes(color = State), size =1/2) +
  labs(title = "The Relationship between Murder Rates & Public School Federal Funds",
       y = "Federal Funds to State Public Schools ($)") +
  scale_x_continuous(name = "All Murder Rates Per 100,000", breaks =seq(0, 21,3)) +
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw() + facet_wrap(vars(State)) +
  theme(legend.position = "none", strip.text = element_text(size = 7, margin=margin()))
```



The relationship between all murder rates and public school federal funds seems to have a slightly negative relationship. Though the relationship seems likely to be non-existent. A few outliers such as California, Texas, and New York show a stronger inverse relationship between all murder rates and public school federal funds.

Figure 6 Figure 6 plots All_Property_Burglary_Rates, Median_Income, and State.

```
# find max burglary rates
summary(rqs$All_Property_Burglary_Rates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   201.7   513.9   690.0   731.4   936.5  1888.8
```

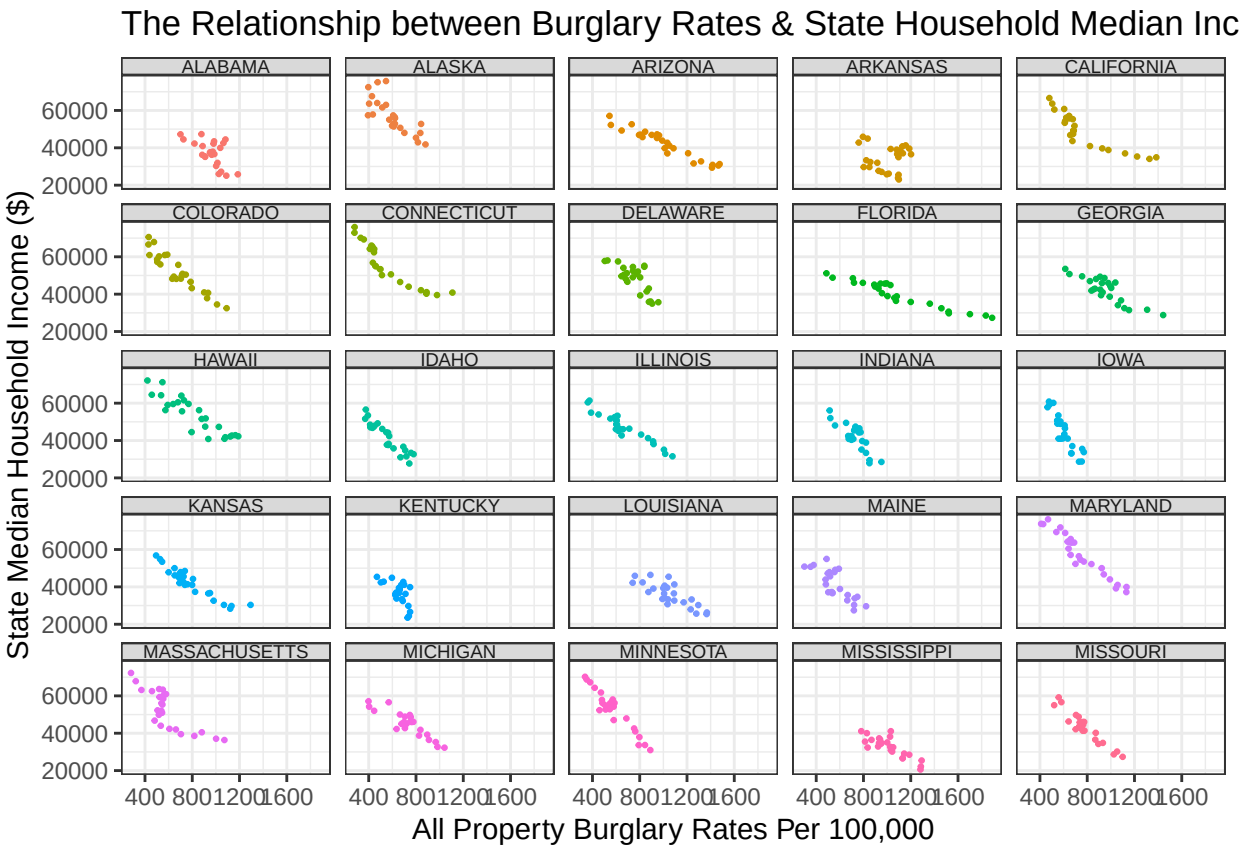
```
# find max murder rates
summary(rqs$All_Property_Burglary_Rates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   201.7   513.9   690.0   731.4   936.5  1888.8
```

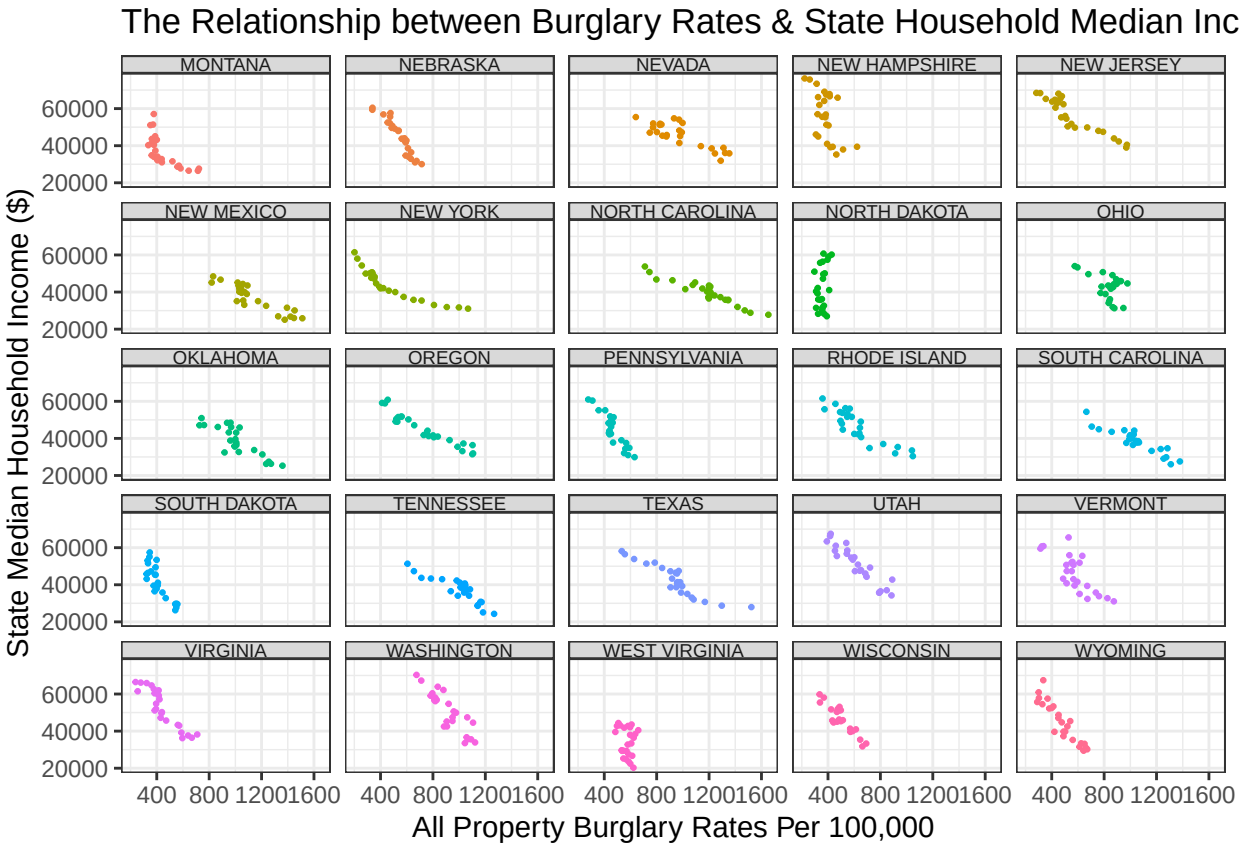
```
# graph 3 variables: y = household median income, x = crime rate (burglary)
BM1 <- rqs %>% arrange(State)
BM1 <- BM1[1:625,]
BM2 <- rqs %>% arrange(State)
BM2 <- BM2[626:1250,]

# graph of first 25 states in alphabetical order
```

```
BM1 %>% ggplot(aes(x = All_Property_Burglary_Rates, y = Median_Income)) +
  geom_point(aes(color = State), size =1/2) +
  labs(title = "The Relationship between Burglary Rates & State Household Median Income",
       y = "State Median Household Income ($)") +
  scale_x_continuous(name = "All Property Burglary Rates Per 100,000",
                     breaks =seq(0, 1890,400)) +
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw() + facet_wrap(vars(State)) +
  theme(legend.position = "none", strip.text = element_text(size = 7, margin=margin()))
```




```
# graph of last 25 states in alphabetical order
BM2 %>% ggplot(aes(x = All_Property_Burglary_Rates,y = Median_Income)) +
  geom_point(aes(color = State), size =1/2) +
  labs(title = "The Relationship between Burglary Rates & State Household Median Income",
       y = "State Median Household Income ($)") +
  scale_x_continuous(name = "All Property Burglary Rates Per 100,000",
                     breaks =seq(0, 1890,400)) +
  # theme(axis.text.x = element_text(angle=90)) +
  theme_bw() + facet_wrap(vars(State)) +
  theme(legend.position = "none",strip.text = element_text(size = 7, margin=margin()))
```



The relationship between all property burglary rates and state household median incomes seems to have a strong negative relationship. Though the relationship seems likely to be non-existent. A few outliers such as Arkansas, Kentucky, and Mississippi show a weaker inverse relationship between all property burglary rates and state household median incomes.

Discussion

Is there a relationship between Federal Funding and Crimes?

Is there a relationship between Median Household Income and Crimes?

Challenges

It was challenging to figure out how to tidy untidy data, especially since there were many special cases we had to fix. This process took us the longest to do. We learned that there's always gonna be caveats with datasets that you need to watch out for, so it is important to have an eye for detail. It was also tricky creating plots because we had to consider how to form them in order to answer our questions. We learned that bouncing off ideas and drawing them out is a good way to figure out which plot is best.

Acknowledgements

We acknowledge the following people:\ **Aileen Li**, worked on:
-Tidying
-Joining/Merging

-Wrangling
-Discussion
Nyah Strickland, worked on:
-Tidying
-Joining/Merging
-Visualizing
-Discussion

References

-Median Household Income by State
-U.S. Education Dataset
-Crime Rate by State Dataset