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:
-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.Murder). The key variables are categorical while all the unique variables are numerical. These are the eight variables required for the project. The following are the links to the datasets:

=======

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(readr)
library(readxl)
```

```r
states_all_extended <- read_csv("states_all_extended.csv")
```

```
## Rows: 1715 Columns: 266
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): PRIMARY_KEY, STATE
## dbl (264): YEAR, ENROLL, TOTAL_REVENUE, FEDERAL_REVENUE, STATE_REVENUE, LOCA...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
state_crime <- read_csv("state_crime.csv")
```

```
## Rows: 3115 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): State
## dbl (20): Year, Data.Population, Data.Rates.Property.All, Data.Rates.Propert...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
h08_1_ <- read_excel("h08 (1).xls")
```

```
## New names:
## * ‘‘ -> ‘...2‘
## * ‘‘ -> ‘...3‘
## * ‘‘ -> ‘...4‘
## * ‘‘ -> ‘...5‘
## * ‘‘ -> ‘...6‘
## * ‘‘ -> ‘...7‘
## * ‘‘ -> ‘...8‘
## * ‘‘ -> ‘...9‘
## * ‘‘ -> ‘...10‘
## * ‘‘ -> ‘...11‘
## * ‘‘ -> ‘...12‘
## * ‘‘ -> ‘...13‘
## * ‘‘ -> ‘...14‘
## * ‘‘ -> ‘...15‘
## * ‘‘ -> ‘...16‘
## * ‘‘ -> ‘...17‘
## * ‘‘ -> ‘...18‘
## * ‘‘ -> ‘...19‘
## * ‘‘ -> ‘...20‘
## * ‘‘ -> ‘...21‘
## * ‘‘ -> ‘...22‘
## * ‘‘ -> ‘...23‘
## * ‘‘ -> ‘...24‘
## * ‘‘ -> ‘...25‘
## * ‘‘ -> ‘...26‘
## * ‘‘ -> ‘...27‘
## * ‘‘ -> ‘...28‘
## * ‘‘ -> ‘...29‘
## * ‘‘ -> ‘...30‘
## * ‘‘ -> ‘...31‘
## * ‘‘ -> ‘...32‘
## * ‘‘ -> ‘...33‘
## * ‘‘ -> ‘...34‘
## * ‘‘ -> ‘...35‘
## * ‘‘ -> ‘...36‘
## * ‘‘ -> ‘...37‘
## * ‘‘ -> ‘...38‘
## * ‘‘ -> ‘...39‘
## * ‘‘ -> ‘...40‘
## * ‘‘ -> ‘...41‘
## * ‘‘ -> ‘...42‘
## * ‘‘ -> ‘...43‘
## * ‘‘ -> ‘...44‘
## * ‘‘ -> ‘...45‘
## * ‘‘ -> ‘...46‘
## * ‘‘ -> ‘...47‘
## * ‘‘ -> ‘...48‘
## * ‘‘ -> ‘...49‘
## * ‘‘ -> ‘...50‘
## * ‘‘ -> ‘...51‘
## * ‘‘ -> ‘...52‘
## * ‘‘ -> ‘...53‘
## * ‘‘ -> ‘...54‘
## * ‘‘ -> ‘...55‘
## * ‘‘ -> ‘...56‘
## * ‘‘ -> ‘...57‘
## * ‘‘ -> ‘...58‘
## * ‘‘ -> ‘...59‘
## * ‘‘ -> ‘...60‘
## * ‘‘ -> ‘...61‘
## * ‘‘ -> ‘...62‘
## * ‘‘ -> ‘...63‘
## * ‘‘ -> ‘...64‘
## * ‘‘ -> ‘...65‘
## * ‘‘ -> ‘...66‘
## * ‘‘ -> ‘...67‘
## * ‘‘ -> ‘...68‘
## * ‘‘ -> ‘...69‘
## * ‘‘ -> ‘...70‘
## * ‘‘ -> ‘...71‘
## * ‘‘ -> ‘...72‘
## * ‘‘ -> ‘...73‘
## * ‘‘ -> ‘...74‘
## * ‘‘ -> ‘...75‘
```

Tidying

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

```r
# Only keep key and unique variables for each dataset

# Eename 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, "_", " "))
```

```r
# 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.

```r
# 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 column 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

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

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

```r
ncol(us_education)
```

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

```r
nrow(crime_rate)
```

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

```r
ncol(crime_rate)
```

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

```r
nrow(med_income)
```

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

```r
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).

```r
# 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 research_questions, 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 research_question 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

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

```
# 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
```

```
summary(rqs$Median_Income)
```

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

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

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

```
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
```

Visualizing

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

PASTE WEBSITES HERE