

# Project 1: Wrangling, Exploration, Visualization

SDS322E

## Data Wrangling, Exploration, Visualization

Shilpi Karan sk46966

**Introduction** The two data sets used in this project are the U.S. State Public School Expenditures dataset and the Violent Crime Rates by US State dataset. The common ID variable that they share is states. Other variables that are in the data sets include murder, assault, urbanpop, and rape in dataset 1 and education, income, young, and urban in dataset2. The variables were acquired by looking at per-capital values and by collecting data from populations of 1,000 or 100,000 people and creating a proportion from that. These data sets and variables were interesting for me because I wanted to know if there is a correlation between education/income and crime rates in America.

```
library(tidyverse)

data1 <- read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/datasets/USArrests.csv")
data2 <- read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/carData/Anscombe.csv")
```

**Tidying: Reshaping** If your datasets are tidy already, demonstrate that you can reshape data with pivot wider/longer here (e.g., untidy and then retidy). Alternatively, it may be easier to wait until the wrangling section so you can reshape your summary statistics. Note here if you are going to do this.

```
data1

## # A tibble: 50 x 5
##   X1      Murder Assault UrbanPop Rape
##   <chr>    <dbl>   <dbl>   <dbl> <dbl>
## 1 Alabama    13.2     236     58  21.2
## 2 Alaska      10     263     48  44.5
## 3 Arizona     8.1     294     80   31
## 4 Arkansas     8.8     190     50  19.5
## 5 California    9     276     91  40.6
## 6 Colorado     7.9     204     78  38.7
## 7 Connecticut   3.3     110     77  11.1
## 8 Delaware     5.9     238     72  15.8
## 9 Florida     15.4     335     80  31.9
## 10 Georgia     17.4     211     60  25.8
## # ... with 40 more rows
```

```
data2

## # A tibble: 51 x 5
##   X1      education income young urban
##   <chr>    <dbl>   <dbl> <dbl> <dbl>
## 1 ME        189    2824   351.  508
## 2 NH        169    3259   346.  564
```

```
## 3 VT      230  3072  348.  322
## 4 MA      168  3835  335.  846
## 5 RI      180  3549  327.  871
## 6 CT      193  4256  341.  774
## 7 NY      261  4151  326.  856
## 8 NJ      214  3954  334.  889
## 9 PA      201  3419  326.  715
## 10 OH     172  3509  354.  753
## # ... with 41 more rows
```

```
# UNTIDY
```

```
data1 <- data1 %>% pivot_wider(names_from = X1, values_from = Murder)
data2 <- data2 %>% pivot_wider(names_from = X1, values_from = education)
```

```
data1
```

```
## # A tibble: 50 x 53
##   Assault UrbanPop Rape Alabama Alaska Arizona Arkansas California Colorado
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 236 58 21.2 13.2 NA NA NA NA NA
## 2 263 48 44.5 NA 10 NA NA NA NA
## 3 294 80 31 NA NA 8.1 NA NA NA
## 4 190 50 19.5 NA NA NA 8.8 NA NA
## 5 276 91 40.6 NA NA NA NA 9 NA
## 6 204 78 38.7 NA NA NA NA NA 7.9
## 7 110 77 11.1 NA NA NA NA NA NA
## 8 238 72 15.8 NA NA NA NA NA NA
## 9 335 80 31.9 NA NA NA NA NA NA
## 10 211 60 25.8 NA NA NA NA NA NA
## # ... with 40 more rows, and 44 more variables: Connecticut <dbl>,
## # Delaware <dbl>, Florida <dbl>, Georgia <dbl>, Hawaii <dbl>, Idaho <dbl>,
## # Illinois <dbl>, Indiana <dbl>, Iowa <dbl>, Kansas <dbl>, Kentucky <dbl>,
## # Louisiana <dbl>, Maine <dbl>, Maryland <dbl>, Massachusetts <dbl>,
## # Michigan <dbl>, Minnesota <dbl>, Mississippi <dbl>, Missouri <dbl>,
## # Montana <dbl>, Nebraska <dbl>, Nevada <dbl>, `New Hampshire` <dbl>, `New
## # Jersey` <dbl>, `New Mexico` <dbl>, `New York` <dbl>, `North
## # Carolina` <dbl>, `North Dakota` <dbl>, Ohio <dbl>, Oklahoma <dbl>,
## # Oregon <dbl>, Pennsylvania <dbl>, `Rhode Island` <dbl>, `South
## # Carolina` <dbl>, `South Dakota` <dbl>, Tennessee <dbl>, Texas <dbl>,
## # Utah <dbl>, Vermont <dbl>, Virginia <dbl>, Washington <dbl>, `West
## # Virginia` <dbl>, Wisconsin <dbl>, Wyoming <dbl>
```

```
data2
```

```
## # A tibble: 51 x 54
##   income young urban ME NH VT MA RI CT NY NJ PA
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2824 351. 508 189 NA NA NA NA NA NA NA NA
## 2 3259 346. 564 NA 169 NA NA NA NA NA NA NA
## 3 3072 348. 322 NA NA 230 NA NA NA NA NA NA NA
## 4 3835 335. 846 NA NA NA 168 NA NA NA NA NA NA NA
## 5 3549 327. 871 NA NA NA NA 180 NA NA NA NA NA NA NA
## 6 4256 341 774 NA NA NA NA NA 193 NA NA NA NA NA NA
## 7 4151 326. 856 NA NA NA NA NA NA 261 NA NA NA NA NA
## 8 3954 334. 889 NA NA NA NA NA NA NA 214 NA NA NA NA
## 9 3419 326. 715 NA NA NA NA NA NA NA NA 201 NA NA NA NA
```

```
## 10 3509 354. 753 NA NA NA NA NA NA NA NA NA
## # ... with 41 more rows, and 42 more variables: OH <dbl>, IN <dbl>, IL <dbl>,
## # MI <dbl>, WI <dbl>, MN <dbl>, IO <dbl>, MO <dbl>, ND <dbl>, SD <dbl>,
## # NE <dbl>, KA <dbl>, DE <dbl>, MD <dbl>, DC <dbl>, VA <dbl>, WV <dbl>,
## # NC <dbl>, SC <dbl>, GA <dbl>, FL <dbl>, KY <dbl>, TN <dbl>, AL <dbl>,
## # MS <dbl>, AR <dbl>, LA <dbl>, OK <dbl>, TX <dbl>, MT <dbl>, ID <dbl>,
## # WY <dbl>, CO <dbl>, NM <dbl>, AZ <dbl>, UT <dbl>, NV <dbl>, WA <dbl>,
## # OR <dbl>, CA <dbl>, AK <dbl>, HI <dbl>
```

*# TIDY*

```
data1 <- data1 %>% pivot_longer(cols = Alabama:Wyoming, names_to = "X1",
  values_to = "Murder", values_drop_na = TRUE)
data2 <- data2 %>% pivot_longer(cols = ME:HI, names_to = "X1",
  values_to = "education", values_drop_na = TRUE)
```

data1

```
## # A tibble: 50 x 5
##   Assault UrbanPop Rape X1      Murder
##   <dbl>    <dbl> <dbl> <chr>    <dbl>
## 1    236      58 21.2 Alabama    13.2
## 2    263      48 44.5 Alaska      10
## 3    294      80 31 Arizona      8.1
## 4    190      50 19.5 Arkansas    8.8
## 5    276      91 40.6 California  9
## 6    204      78 38.7 Colorado    7.9
## 7    110      77 11.1 Connecticut 3.3
## 8    238      72 15.8 Delaware    5.9
## 9    335      80 31.9 Florida    15.4
## 10   211      60 25.8 Georgia     17.4
## # ... with 40 more rows
```

data2

```
## # A tibble: 51 x 5
##   income young urban X1      education
##   <dbl> <dbl> <dbl> <chr>    <dbl>
## 1  2824  351.  508 ME      189
## 2  3259  346.  564 NH      169
## 3  3072  348.  322 VT      230
## 4  3835  335.  846 MA      168
## 5  3549  327.  871 RI      180
## 6  4256  341  774 CT      193
## 7  4151  326.  856 NY      261
## 8  3954  334.  889 NJ      214
## 9  3419  326.  715 PA      201
## 10  3509  354.  753 OH      172
## # ... with 41 more rows
```

```
data1 <- data1 %>% relocate(X1, .before = Assault)
data2 <- data2 %>% relocate(X1, .before = income)
```

data1

```
## # A tibble: 50 x 5
##   X1      Assault UrbanPop Rape Murder
##   <chr>    <dbl>    <dbl> <dbl> <dbl>
```

```
## 1 Alabama      236      58 21.2 13.2
## 2 Alaska       263      48 44.5 10
## 3 Arizona      294      80 31    8.1
## 4 Arkansas     190      50 19.5 8.8
## 5 California   276      91 40.6 9
## 6 Colorado     204      78 38.7 7.9
## 7 Connecticut  110      77 11.1 3.3
## 8 Delaware     238      72 15.8 5.9
## 9 Florida      335      80 31.9 15.4
## 10 Georgia     211      60 25.8 17.4
## # ... with 40 more rows
```

```
data2
```

```
## # A tibble: 51 x 5
##   X1      income young urban education
##   <chr>   <dbl> <dbl> <dbl>      <dbl>
## 1 ME      2824  351.  508      189
## 2 NH      3259  346.  564      169
## 3 VT      3072  348.  322      230
## 4 MA      3835  335.  846      168
## 5 RI      3549  327.  871      180
## 6 CT      4256  341   774      193
## 7 NY      4151  326.  856      261
## 8 NJ      3954  334.  889      214
## 9 PA      3419  326.  715      201
## 10 OH     3509  354.  753      172
## # ... with 41 more rows
```

```
library(dplyr)
```

```
data2[1, 1] <- "Alaska"
data2[2, 1] <- "Alabama"
data2[3, 1] <- "Arkansas"
data2[4, 1] <- "Arizona"
data2[5, 1] <- "California"
data2[6, 1] <- "Colorado"
data2[7, 1] <- "Connecticut"
data2[8, 1] <- "Delaware"
data2[9, 1] <- "Florida"
data2[10, 1] <- "Georgia"
data2[11, 1] <- "Hawaii"
data2[12, 1] <- "Idaho"
data2[13, 1] <- "Illinois"
data2[14, 1] <- "Indiana"
data2[15, 1] <- "Iowa"
data2[16, 1] <- "Kansas"
data2[17, 1] <- "Kentucky"
data2[18, 1] <- "Louisiana"
data2[19, 1] <- "Massachusetts"
data2[20, 1] <- "Maryland"
data2[21, 1] <- "Maine"
data2[22, 1] <- "Michigan"
```

```

data2[23, 1] <- "Minnesota"
data2[24, 1] <- "Missouri"
data2[25, 1] <- "Mississippi"
data2[26, 1] <- "Montana"
data2[27, 1] <- "North Carolina"
data2[28, 1] <- "North Dakota"
data2[29, 1] <- "Nebraska"
data2[30, 1] <- "New Hampshire"
data2[31, 1] <- "New Jersey"
data2[32, 1] <- "New Mexico"
data2[33, 1] <- "Nevada"
data2[34, 1] <- "New York"
data2[35, 1] <- "Ohio"
data2[36, 1] <- "Oklahoma"
data2[37, 1] <- "Oregon"
data2[38, 1] <- "Pennsylvania"
data2[39, 1] <- "Rhode Island"
data2[40, 1] <- "South Carolina"
data2[41, 1] <- "South Dakota"
data2[42, 1] <- "Tennessee"
data2[43, 1] <- "Texas"
data2[44, 1] <- "Utah"
data2[45, 1] <- "Virginia"
data2[46, 1] <- "Vermont"
data2[47, 1] <- "Washington"
data2[48, 1] <- "Wisconsin"
data2[49, 1] <- "West Virginia"
data2[50, 1] <- "Wyoming"

```

```

data2 <- data2 %>% arrange(data2)
data2

```

## Joining/Merging

```

## # A tibble: 51 x 5
##   X1          income young urban education
##   <chr>      <dbl> <dbl> <dbl>      <dbl>
## 1 Alabama    3259  346.  564        169
## 2 Alaska     2824  351.  508        189
## 3 Arizona    3835  335.  846        168
## 4 Arkansas   3072  348.  322        230
## 5 California 3549  327.  871        180
## 6 Colorado   4256  341.  774        193
## 7 Connecticut 4151  326.  856        261
## 8 Delaware   3954  334.  889        214
## 9 Florida    3419  326.  715        201
## 10 Georgia   3509  354.  753        172
## # ... with 41 more rows

```

```

data3 <- full_join(data1, data2)
data3

```

```

## # A tibble: 51 x 9
##   X1          Assault UrbanPop Rape Murder income young urban education
##   <chr>      <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>      <dbl>

```

```
## 1 Alabama      236      58 21.2 13.2 3259 346. 564      169
## 2 Alaska       263      48 44.5 10   2824 351. 508      189
## 3 Arizona      294      80 31   8.1 3835 335. 846      168
## 4 Arkansas     190      50 19.5 8.8 3072 348. 322      230
## 5 California   276      91 40.6 9   3549 327. 871      180
## 6 Colorado     204      78 38.7 7.9 4256 341  774      193
## 7 Connecticut  110      77 11.1 3.3 4151 326. 856      261
## 8 Delaware     238      72 15.8 5.9 3954 334. 889      214
## 9 Florida      335      80 31.9 15.4 3419 326. 715      201
## 10 Georgia     211      60 25.8 17.4 3509 354. 753      172
## # ... with 41 more rows
```

```
# data1 <- data1 %>% rename(States = X1) data2 <- data2 %>%
# rename(States = X1)
data1
```

```
## # A tibble: 50 x 5
##   X1      Assault UrbanPop Rape Murder
##   <chr>      <dbl>    <dbl> <dbl> <dbl>
## 1 Alabama      236      58 21.2 13.2
## 2 Alaska       263      48 44.5 10
## 3 Arizona      294      80 31   8.1
## 4 Arkansas     190      50 19.5 8.8
## 5 California   276      91 40.6 9
## 6 Colorado     204      78 38.7 7.9
## 7 Connecticut  110      77 11.1 3.3
## 8 Delaware     238      72 15.8 5.9
## 9 Florida      335      80 31.9 15.4
## 10 Georgia     211      60 25.8 17.4
## # ... with 40 more rows
```

```
data2
```

```
## # A tibble: 51 x 5
##   X1      income young urban education
##   <chr>      <dbl> <dbl> <dbl>    <dbl>
## 1 Alabama      3259 346. 564      169
## 2 Alaska       2824 351. 508      189
## 3 Arizona      3835 335. 846      168
## 4 Arkansas     3072 348. 322      230
## 5 California   3549 327. 871      180
## 6 Colorado     4256 341  774      193
## 7 Connecticut  4151 326. 856      261
## 8 Delaware     3954 334. 889      214
## 9 Florida      3419 326. 715      201
## 10 Georgia     3509 354. 753      172
## # ... with 41 more rows
```

```
dim(data1)
```

```
## [1] 50 5
```

```
dim(data2)
```

```
## [1] 51 5
```

```
dim(data3)
```

```
## [1] 51 9
```

```
colnames(data1)
```

```
## [1] "X1" "Assault" "UrbanPop" "Rape" "Murder"
```

```
colnames(data2)
```

```
## [1] "X1" "income" "young" "urban" "education"
```

Datasets 1 and 2 were full joined. I used full join because both datasets have the same matching rows so it would make no difference and thus be of no use to do inner, left, or right join. There are 50 observations/rows in each dataset. The ID that the datasets have in common is states. The unique IDs in dataset 1 that are not in dataset 2 are murder, assault, urbanpop, and rape. The other IDs are unique to dataset 2, and they are education, income, young, and urban. The size of the joined dataset is larger than the individual datasets. It has 9 variables/columns while the individual datasets had 5 columns/variables each. There were no observations dropped, and so there is also no problem associated with it.

```
data3 %>% arrange(income)
```

## Wrangling

```
## # A tibble: 51 x 9
```

```
##   X1          Assault UrbanPop Rape Murder income young urban education
##   <chr>         <dbl>    <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>    <dbl>
## 1 New York      254        86  26.1  11.1   2081  385.  445     130
## 2 Ohio          120        75  21.4   7.3   2322  352.  500     134
## 3 Nevada        252        81  46    12.2   2337  362.  584     112
## 4 North Dakota   45        44   7.3   0.8   2380  377.  476     149
## 5 Montana        109        53  16.4   6     2470  329.  390     149
## 6 New Mexico     285        70  32.1  11.4   2579  343.  588     137
## 7 Oklahoma       151        68  20     6.6   2634  390.  661     162
## 8 New Jersey     159        89  18.8   7.4   2645  349.  523     140
## 9 Texas          201        80  25.5  12.7   2651  422.  698     227
## 10 North Carolina 337        45  16.1  13     2664  354.  450     155
## # ... with 41 more rows
```

```
data3 %>% filter(str_detect(Rape, "17.4"))
```

```
## # A tibble: 0 x 9
```

```
## # ... with 9 variables: X1 <chr>, Assault <dbl>, UrbanPop <dbl>, Rape <dbl>,
## #   Murder <dbl>, income <dbl>, young <dbl>, urban <dbl>, education <dbl>
```

```
data3 %>% filter(urban >= 500)
```

```
## # A tibble: 43 x 9
```

```
##   X1          Assault UrbanPop Rape Murder income young urban education
##   <chr>         <dbl>    <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>    <dbl>
## 1 Alabama      236        58  21.2  13.2   3259  346.  564     169
## 2 Alaska        263        48  44.5  10     2824  351.  508     189
## 3 Arizona       294        80  31     8.1   3835  335.  846     168
## 4 California    276        91  40.6   9     3549  327.  871     180
## 5 Colorado      204        78  38.7   7.9   4256  341.  774     193
## 6 Connecticut   110        77  11.1   3.3   4151  326.  856     261
## 7 Delaware      238        72  15.8   5.9   3954  334.  889     214
## 8 Florida       335        80  31.9  15.4   3419  326.  715     201
## 9 Georgia       211        60  25.8  17.4   3509  354.  753     172
```

```
## 10 Hawaii          46      83 20.2    5.3 3412 359.  649      194
## # ... with 33 more rows
```

```
data3 %>% select(X1, Murder, income)
```

```
## # A tibble: 51 x 3
##   X1      Murder income
##   <chr>    <dbl> <dbl>
## 1 Alabama    13.2  3259
## 2 Alaska     10    2824
## 3 Arizona     8.1  3835
## 4 Arkansas     8.8  3072
## 5 California     9    3549
## 6 Colorado     7.9  4256
## 7 Connecticut    3.3  4151
## 8 Delaware     5.9  3954
## 9 Florida    15.4  3419
## 10 Georgia    17.4  3509
## # ... with 41 more rows
```

```
data3 %>% mutate(ratio = Murder/income)
```

```
## # A tibble: 51 x 10
##   X1      Assault UrbanPop Rape Murder income young urban education ratio
##   <chr>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl>
## 1 Alabama    236      58 21.2  13.2  3259 346.  564      169 4.05e-3
## 2 Alaska    263      48 44.5   10    2824 351.  508      189 3.54e-3
## 3 Arizona    294      80 31     8.1  3835 335.  846      168 2.11e-3
## 4 Arkansas   190      50 19.5   8.8  3072 348.  322      230 2.86e-3
## 5 California 276      91 40.6   9    3549 327.  871      180 2.54e-3
## 6 Colorado   204      78 38.7   7.9  4256 341.  774      193 1.86e-3
## 7 Connectic~ 110      77 11.1   3.3  4151 326.  856      261 7.95e-4
## 8 Delaware   238      72 15.8   5.9  3954 334.  889      214 1.49e-3
## 9 Florida    335      80 31.9  15.4  3419 326.  715      201 4.50e-3
## 10 Georgia   211      60 25.8  17.4  3509 354.  753      172 4.96e-3
## # ... with 41 more rows
```

```
data3$lh_UrbanPop <- as.factor(ifelse(data3$UrbanPop < 50, "low",
"high"))
data3
```

```
## # A tibble: 51 x 10
##   X1      Assault UrbanPop Rape Murder income young urban education lh_UrbanPop
##   <chr>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <fct>
## 1 Alaba~    236      58 21.2  13.2  3259 346.  564      169 high
## 2 Alaska    263      48 44.5   10    2824 351.  508      189 low
## 3 Arizo~    294      80 31     8.1  3835 335.  846      168 high
## 4 Arkan~    190      50 19.5   8.8  3072 348.  322      230 high
## 5 Calif~    276      91 40.6   9    3549 327.  871      180 high
## 6 Color~    204      78 38.7   7.9  4256 341.  774      193 high
## 7 Conne~    110      77 11.1   3.3  4151 326.  856      261 high
## 8 Delaw~    238      72 15.8   5.9  3954 334.  889      214 high
## 9 Flori~    335      80 31.9  15.4  3419 326.  715      201 high
## 10 Georg~    211      60 25.8  17.4  3509 354.  753      172 high
## # ... with 41 more rows
```



```
data3 %>% group_by(lh_UrbanPop) %>% summarize(counts = n())
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop counts
##   <fct>         <int>
## 1 high           42
## 2 low            8
## 3 <NA>           1
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(education,
  na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `mean(education, na.rm = T)`
##   <fct>         <dbl>
## 1 high           196.
## 2 low           197.
## 3 <NA>          212
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(education, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(education, na.rm = T)`
##   <fct>         <dbl>
## 1 high           47.9
## 2 low           43.7
## 3 <NA>           NA
```

```
data3 %>% summarize(max(education, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(education, na.rm = T)`
##   <dbl>
## 1           372
```

```
data3 %>% summarize(min(education, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(education, na.rm = T)`
##   <dbl>
## 1           112
```

```
data3 %>% summarize(median(education, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(education, na.rm = T)`
##   <dbl>
## 1           192
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(Murder, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `mean(Murder, na.rm = T)`
##   <fct>         <dbl>
## 1 high           7.7
## 2 low           8.25
## 3 <NA>          NaN
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(Murder, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(Murder, na.rm = T)`
##   <fct>      <dbl>
## 1 high      4.08
## 2 low       5.90
## 3 <NA>      NA
```

```
data3 %>% summarize(max(Murder, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(Murder, na.rm = T)`
##   <dbl>
## 1      17.4
```

```
data3 %>% summarize(min(Murder, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(Murder, na.rm = T)`
##   <dbl>
## 1      0.8
```

```
data3 %>% summarize(median(Murder, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(Murder, na.rm = T)`
##   <dbl>
## 1      7.25
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(income, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `mean(income, na.rm = T)`
##   <fct>      <dbl>
## 1 high     3244.
## 2 low     3090.
## 3 <NA>     3513
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(income, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(income, na.rm = T)`
##   <fct>      <dbl>
## 1 high      562.
## 2 low      594.
## 3 <NA>      NA
```

```
data3 %>% summarize(max(income, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(income, na.rm = T)`
##   <dbl>
## 1      4425
```

```
data3 %>% summarize(min(income, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(income, na.rm = T)`
```

```
##                               <dbl>
## 1                             2081

data3 %>% summarize(median(income, na.rm = T))

## # A tibble: 1 x 1
##   `median(income, na.rm = T)`
##                               <dbl>
## 1                             3257

data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(UrbanPop,
  na.rm = T))

## # A tibble: 3 x 2
##   lh_UrbanPop `mean(UrbanPop, na.rm = T)`
##   <fct>                               <dbl>
## 1 high                               69.8
## 2 low                                43.1
## 3 <NA>                               NaN

data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(UrbanPop, na.rm = T))

## # A tibble: 3 x 2
##   lh_UrbanPop `sd(UrbanPop, na.rm = T)`
##   <fct>                               <dbl>
## 1 high                               11.4
## 2 low                                5.30
## 3 <NA>                               NA

data3 %>% summarize(max(UrbanPop, na.rm = T))

## # A tibble: 1 x 1
##   `max(UrbanPop, na.rm = T)`
##                               <dbl>
## 1                             91

data3 %>% summarize(min(UrbanPop, na.rm = T))

## # A tibble: 1 x 1
##   `min(UrbanPop, na.rm = T)`
##                               <dbl>
## 1                             32

data3 %>% summarize(median(UrbanPop, na.rm = T))

## # A tibble: 1 x 1
##   `median(UrbanPop, na.rm = T)`
##                               <dbl>
## 1                             66

data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(Assault, na.rm = T))

## # A tibble: 3 x 2
##   lh_UrbanPop `mean(Assault, na.rm = T)`
##   <fct>                               <dbl>
## 1 high                               170
## 2 low                                175.
## 3 <NA>                               NaN
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(Assault, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(Assault, na.rm = T)`
##   <fct>      <dbl>
## 1 high      76.3
## 2 low      121.
## 3 <NA>      NA
```

```
data3 %>% summarize(max(Assault, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(Assault, na.rm = T)`
##   <dbl>
## 1      337
```

```
data3 %>% summarize(min(Assault, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(Assault, na.rm = T)`
##   <dbl>
## 1      45
```

```
data3 %>% summarize(median(Assault, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(Assault, na.rm = T)`
##   <dbl>
## 1      159
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(young, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `mean(young, na.rm = T)`
##   <fct>      <dbl>
## 1 high      358.
## 2 low      363.
## 3 <NA>      383.
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(young, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(young, na.rm = T)`
##   <fct>      <dbl>
## 1 high      25.5
## 2 low      13.4
## 3 <NA>      NA
```

```
data3 %>% summarize(max(young, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(young, na.rm = T)`
##   <dbl>
## 1      440.
```

```
data3 %>% summarize(min(young, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(young, na.rm = T)`
```

```
##                               <dbl>
## 1                             326.
data3 %>% summarize(median(young, na.rm = T))

## # A tibble: 1 x 1
##   `median(young, na.rm = T)`
##                               <dbl>
## 1                             354.
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(Rape, na.rm = T))

## # A tibble: 3 x 2
##   lh_UrbanPop `mean(Rape, na.rm = T)`
##   <fct>                <dbl>
## 1 high                21.9
## 2 low                 17.6
## 3 <NA>                 NaN
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(Rape, na.rm = T))

## # A tibble: 3 x 2
##   lh_UrbanPop `sd(Rape, na.rm = T)`
##   <fct>                <dbl>
## 1 high                8.81
## 2 low                 11.9
## 3 <NA>                 NA
data3 %>% summarize(max(Rape, na.rm = T))

## # A tibble: 1 x 1
##   `max(Rape, na.rm = T)`
##                               <dbl>
## 1                             46
data3 %>% summarize(min(Rape, na.rm = T))

## # A tibble: 1 x 1
##   `min(Rape, na.rm = T)`
##                               <dbl>
## 1                             7.3
data3 %>% summarize(median(Rape, na.rm = T))

## # A tibble: 1 x 1
##   `median(Rape, na.rm = T)`
##                               <dbl>
## 1                             20.1
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(urban, na.rm = T))

## # A tibble: 3 x 2
##   lh_UrbanPop `mean(urban, na.rm = T)`
##   <fct>                <dbl>
## 1 high                670.
## 2 low                 616.
## 3 <NA>                 831
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(urban, na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(urban, na.rm = T)`
##   <fct>      <dbl>
## 1 high      149.
## 2 low       164.
## 3 <NA>      NA
```

```
data3 %>% summarize(max(urban, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(urban, na.rm = T)`
##   <dbl>
## 1      1000
```

```
data3 %>% summarize(min(urban, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(urban, na.rm = T)`
##   <dbl>
## 1      322
```

```
data3 %>% summarize(median(urban, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(urban, na.rm = T)`
##   <dbl>
## 1      664
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(mean(education/income,
  na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `mean(education/income, na.rm = T)`
##   <fct>      <dbl>
## 1 high      0.0605
## 2 low       0.0638
## 3 <NA>      0.0603
```

```
data3 %>% group_by(lh_UrbanPop) %>% summarize(sd(education/income,
  na.rm = T))
```

```
## # A tibble: 3 x 2
##   lh_UrbanPop `sd(education/income, na.rm = T)`
##   <fct>      <dbl>
## 1 high      0.0109
## 2 low       0.00608
## 3 <NA>      NA
```

```
data3 %>% summarize(max(education/income, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `max(education/income, na.rm = T)`
##   <dbl>
## 1      0.0897
```

```
data3 %>% summarize(min(education/income, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `min(education/income, na.rm = T)`
##   <dbl>
```

```
## 1 0.0438
data3 %>% summarize(median(education/income, na.rm = T))

## # A tibble: 1 x 1
##   `median(education/income, na.rm = T)`
##                                     <dbl>
## 1 0.0599

percent_decimal <- function(UrbanPop) {
  DecUrbanPop <- (UrbanPop/100)
  return(DecUrbanPop)
}

data3 %>% summarize(percent_decimal(max(UrbanPop, na.rm = T)))

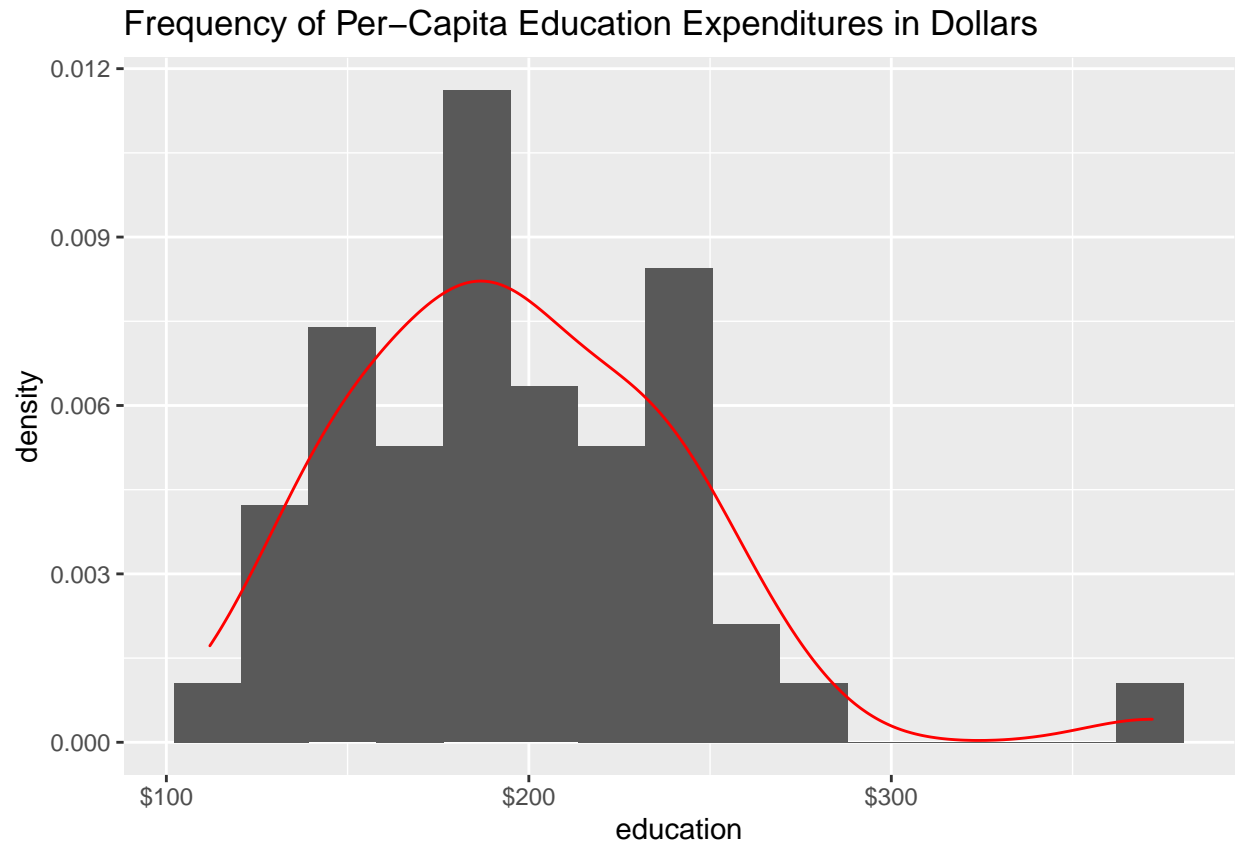
## # A tibble: 1 x 1
##   `percent_decimal(max(UrbanPop, na.rm = T))`
##                                     <dbl>
## 1 0.91

# gt_tbl <- gt(data3 %>% group_by(lh_UrbanPop) %>%
# summarize(counts=n())) gt_tbl
```

To start off, the data was arranged based on income. Then, it was filtered for certain variables, selected for other variables, and mutated for certain values. All of this was done to better understand the relationship (and if there was one or not) between crime rates and income/education. A new categorical variable was created that sorted the data into high and low urbanpop percentages. Later, the data was grouped by the new categorical variable that was created and it was also used in the summaries in which, mean, median, max, min, and standard deviation values were determined for each variable. The counts for the low and high urbanpop was also determined. One table was also styled with a gt package. A new function was also created to help with making sense of the data more. Byfar, the most interesting finding was that a larger than expected amount of the U.S. population lived in urban areas). Another interesting finding was that, there is one state where .91 out of 1 of the population lives in Urban areas (noted as a decimal), and this is indicated by the function that was created.

```
ggplot(data3, aes(x = education)) + geom_histogram(aes(y = ..density..),
  bins = 15) + geom_density(color = "red") + theme_grey() +
  scale_x_continuous(labels = scales::dollar) + ggtitle("Frequency of Per-Capita Education Expenditure")
```

## Visualizing

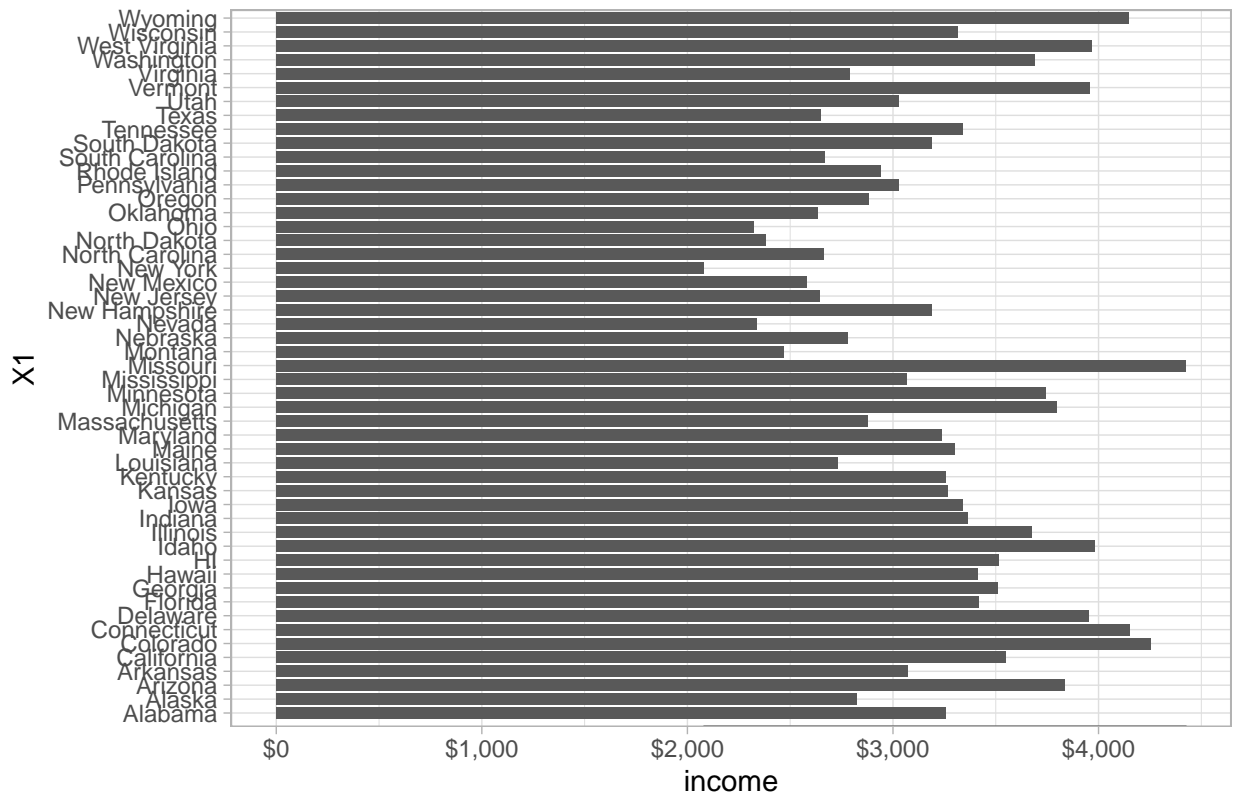


The plot depicts the amount of money that was spent per-capita on education in dollars. The relationships/trends that are apparent from this histogram and density line is that most people did not spend as much on education and that the majority spent around \$180. The plot shows mainly a normal distribution bell curve with a potential outlier to the far right. Overall, this plot indicates that the majority of the population spends similar amounts of money towards education relative to one another.

```
ggplot(data3, aes(x = income)) + geom_bar(aes(y = X1), stat = "summary",
width = 0.8) + geom_density() + theme_light() + scale_x_continuous(labels = scales::dollar) +
ggtitle("Per-capita Income in Each U.S. State in Dollars")
```

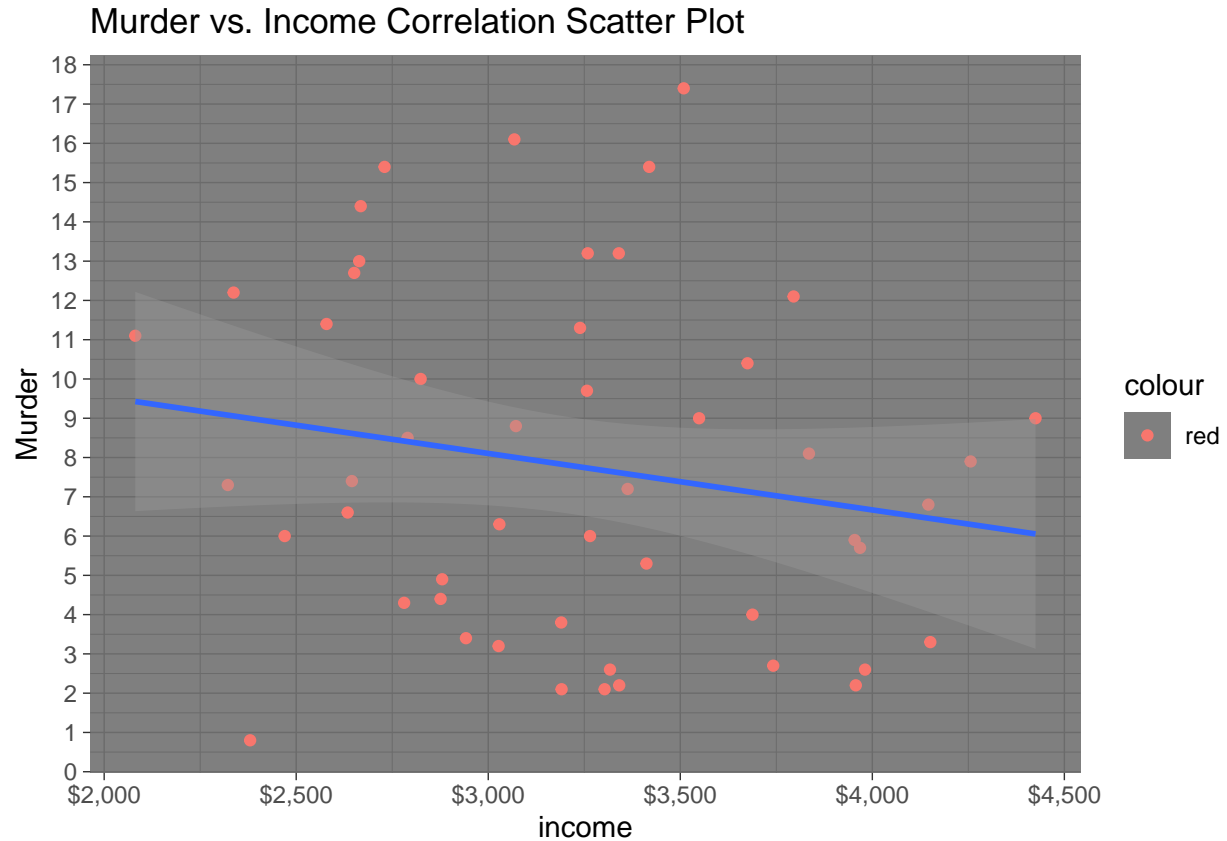


Per-capita Income in Each U.S. State in Dollars



The barplot shown above depicts income per-capita, in dollars, in each U.S. state. There is no apparent trend or relationship shown from this plot. It is only apparent that some states have a much higher per-capita income than other states. This just shows the variations of incomes between different states.

```
ggplot(data = data3, aes(x = income, y = Murder)) + geom_point(aes(color = "red")) +
  geom_smooth(method = "lm") + theme_dark() + scale_x_continuous(labels = scales::dollar) +
  scale_y_continuous(breaks = seq(0, 18, 1)) + ggtitle("Murder vs. Income Correlation Scatter Plot")
```



The scatterplot above shows the correlation between murder and income. Based on the plot, it is apparent that there is no correlation between the two variables. There is no obvious relationship as the values for income and murder for each state is mainly scattered. The trendline also shows that there is no positive or negative linear relationship.

**Concluding Remarks** We cannot conclude anything from this data in regards to the relation between education/income values and crime rates. There does not seem to be an apparent relationship from the data that was collected, so there cannot be a conclusion or generalization made from it.