Project 2

Mathew Katz

2022-10-08

```
library(tidyr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
```

Data for dataset number one is captured from the 2017 American Community Survey using the tidycensus package. We only want to look at the name, whether the dollar amount is income or rent, and the estimated dollar amount.

```
df1=us_rent_income
df2=df1[c("NAME","variable","estimate")]
head(df2)
```

```
## # A tibble: 6 x 3
##   NAME    variable estimate
##   <chr>   <chr>       <dbl>
## 1 Alabama income      24476
## 2 Alabama rent          747
## 3 Alaska  income      32940
## 4 Alaska  rent         1200
## 5 Arizona income      27517
## 6 Arizona rent          972
```

Let's "widen" the data; increasing the number of columns and decreasing the number of rows.

```r
df=pivot_wider(df2,names_from=variable,values_from = estimate)
head(df)
```

```
## # A tibble: 6 x 3
##   NAME       income  rent
##   <chr>       <dbl> <dbl>
## 1 Alabama     24476   747
## 2 Alaska      32940  1200
## 3 Arizona     27517   972
## 4 Arkansas    23789   709
## 5 California  29454  1358
## 6 Colorado    32401  1125
```

Let's look at the US territories sorted by income (high to low.)

```r
income_sorted<- df[order(df$income, decreasing = TRUE),]
income_sorted
```

```
## # A tibble: 52 x 3
##    NAME                 income  rent
##    <chr>                 <dbl> <dbl>
##  1 District of Columbia  43198  1424
##  2 Maryland              37147  1311
##  3 Connecticut           35326  1123
##  4 New Jersey            35075  1249
##  5 Massachusetts         34498  1173
##  6 New Hampshire         33172  1052
##  7 Alaska                32940  1200
##  8 Minnesota             32734   906
##  9 Virginia              32545  1166
## 10 Hawaii                32453  1507
## # ... with 42 more rows
```

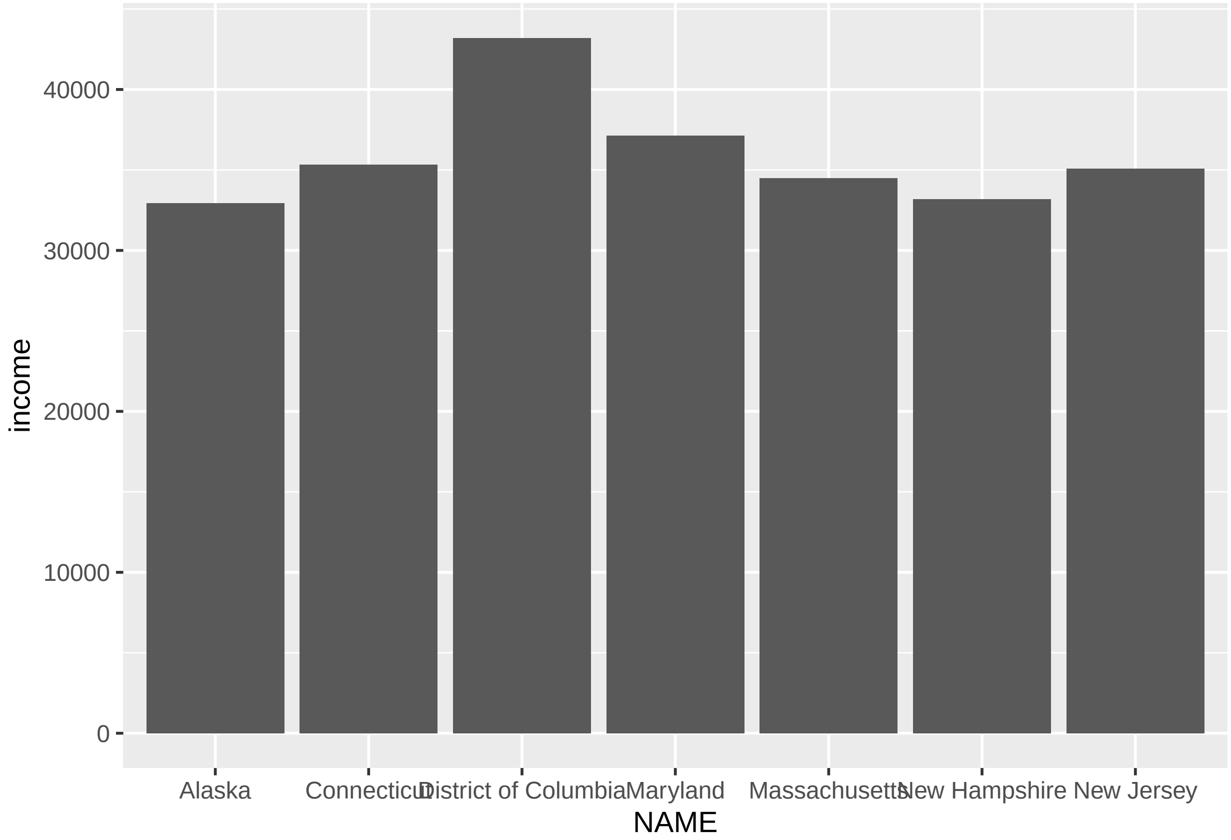
Let's look at the US territories sorted by rent (high to low.)

```r
rent_sorted<- df[order(df$rent, decreasing = TRUE),]
rent_sorted
```

```
## # A tibble: 52 x 3
##    NAME                 income  rent
##    <chr>                 <dbl> <dbl>
##  1 Hawaii                32453  1507
##  2 District of Columbia  43198  1424
##  3 California            29454  1358
##  4 Maryland              37147  1311
##  5 New Jersey            35075  1249
##  6 Alaska                32940  1200
##  7 New York              31057  1194
##  8 Massachusetts         34498  1173
##  9 Virginia              32545  1166
## 10 Colorado              32401  1125
## # ... with 42 more rows
```

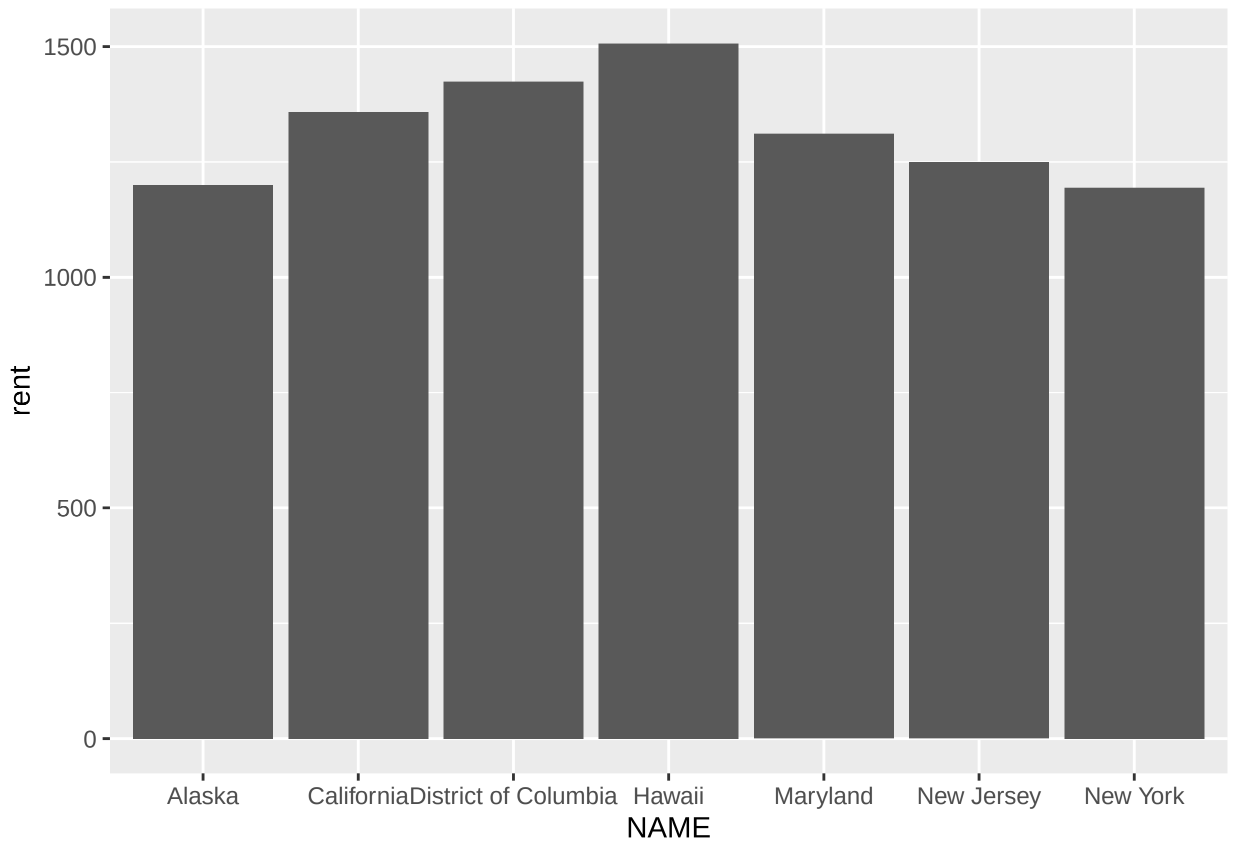
This bar plot will show the seven highest income territories in the US.

```
library(ggplot2)
i<-ggplot(data=head(income_sorted, n=7), aes(x=NAME, y=income)) +
  geom_bar(stat="identity")
i
```



This bar plot will show the seven highest rent territories in the US.

```
r<-ggplot(data=head(rent_sorted, n=7), aes(x=NAME, y=rent)) +
  geom_bar(stat="identity")
r
```



Data for dataset number two looks at median weekly earnings of full-time wage and salary workers by detailed occupation and sex.

```
work = read_csv('work.csv')
```

```
## Rows: 558 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Occupation, All_weekly, M_weekly, F_weekly
## dbl (3): All_workers, M_workers, F_workers
##
## 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.
```

```
head(work)
```

```
## # A tibble: 6 x 7
##   Occupation                          All_w~1 All_w~2 M_wor~3 M_wee~4 F_wor~5 F_wee~6
##   <chr>                                 <dbl> <chr>     <dbl> <chr>     <dbl> <chr>
## 1 ALL OCCUPATIONS                      109080 809       60746 895       48334 726
## 2 MANAGEMENT                            12480 1351       7332 1486       5147 1139
## 3 Chief executives                       1046 2041        763 2251        283 1836
## 4 General and operations manage~          823 1260        621 1347        202 1002
## 5 Legislators                               8 Na            5 Na            4 Na
## 6 Advertising and promotions ma~           55 1050         29 Na           26 Na
## # ... with abbreviated variable names 1: All_workers, 2: All_weekly,
## #   3: M_workers, 4: M_weekly, 5: F_workers, 6: F_weekly
```

Let's "gather" a key-value pair across multiple columns and also separate the data frame into multiple columns.

```
work %>%
    gather(key, value, 2:7) %>%
    separate(key, into=c("gender", "class"), sep="_") -> work1
head(work1)
```

```
## # A tibble: 6 x 4
##   Occupation                          gender class   value
##   <chr>                               <chr>  <chr>   <chr>
## 1 ALL OCCUPATIONS                     All    workers 109080
## 2 MANAGEMENT                          All    workers 12480
## 3 Chief executives                    All    workers 1046
## 4 General and operations managers     All    workers 823
## 5 Legislators                         All    workers 8
## 6 Advertising and promotions managers All    workers 55
```

Let's change the dollar amounts from characters to numbers and get rid of NAs.

```
work1 %>% mutate(value=as.numeric(value)) %>% na.omit() -> work1
```

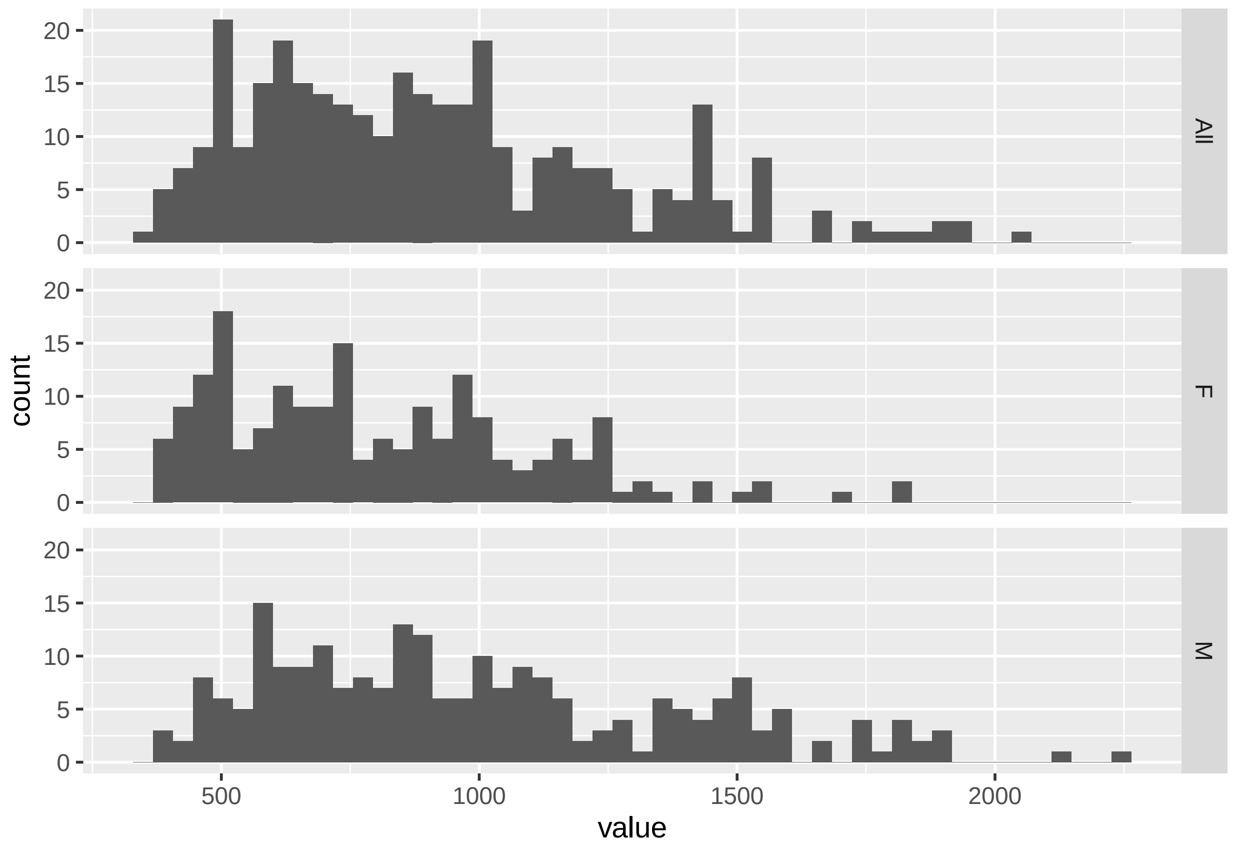
```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
head(work1)
```

```
## # A tibble: 6 x 4
##   Occupation                          gender class    value
##   <chr>                               <chr>  <chr>    <dbl>
## 1 ALL OCCUPATIONS                     All    workers 109080
## 2 MANAGEMENT                          All    workers  12480
## 3 Chief executives                    All    workers   1046
## 4 General and operations managers     All    workers    823
## 5 Legislators                         All    workers      8
## 6 Advertising and promotions managers All    workers     55
```

Histogram of Male, Female, and Total weekly income:

```
work1 %>%
    filter(class=='weekly') %>%
    ggplot(aes(x=value)) +
    geom_histogram(bins=50) +
    facet_grid(gender ~ .)
```



Male, Female, and Total median income:

```
work1 %>%
    filter(class == 'weekly') %>%
    group_by(gender) %>%
    summarize(median(value))
```

```
## # A tibble: 3 x 2
##   gender `median(value)`
##   <chr>            <dbl>
## 1 All               856
## 2 F                 736
## 3 M                 916.
```

Verification of Male and Female median income:

```
median(na.omit(as.numeric(work$M_weekly)))
```

```
## Warning in na.omit(as.numeric(work$M_weekly)): NAs introduced by coercion
```

```
## [1] 915.5
```

```
median(na.omit(as.numeric(work$F_weekly)))
```

```
## Warning in na.omit(as.numeric(work$F_weekly)): NAs introduced by coercion
## [1] 736
```

Data for dataset number three is a database of all of the mass shootings in the US from 1966-2016.

```
mass_shootings <- read_csv('mass_shootings.csv')
```

```
## Rows: 347 Columns: 48
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (41): Title, Location, City, State, Description, Date, Day of Week, Date...
## dbl  (7): CaseID, Latitude, Longitude, Number of Victim Fatalities, Total Nu...
##
## 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.
```

```
head(mass_shootings)
```

```
## # A tibble: 6 x 48
##   CaseID Title       Locat~1 City  State Latit~2 Longi~3 Numbe~4 Total~5 Numbe~6
##    <dbl> <chr>       <chr>   <chr> <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1      1 University~ Austin~ Aust~ Texas    30.2   -97.8      16      17      32
## 2      2 Rose-Mar C~ Mesa, ~ Mesa  Ariz~    33.4  -112.        5       5       1
## 3      3 New Orlean~ New Or~ New ~ Loui~    30.1   -89.9       9      10      13
## 4      4 Clara Bart~ Chicag~ Chic~ Illi~    41.8   -87.7       1       1       3
## 5      5 Olean High~ Olean,~ Olean New ~    42.1   -78.4       3       3       7
## 6      6 Los Angele~ Los An~ Los ~ Cali~    34.2  -119.        1       1       7
## # ... with 38 more variables: ‘Total Number of Victims‘ <dbl>,
## #   Description <chr>, Date <chr>, ‘Day of Week‘ <chr>,
## #   ‘Date - Detailed‘ <chr>, ‘Shooter Name‘ <chr>, ‘Shooter Age(s)‘ <chr>,
## #   ‘Average Shooter Age‘ <chr>, ‘Shooter Sex‘ <chr>, ‘Shooter Race‘ <chr>,
## #   ‘Type of Gun - Detailed‘ <chr>, ‘Type of Gun - General‘ <chr>,
## #   ‘Number of Shotguns‘ <chr>, ‘Number of Rifles‘ <chr>,
## #   ‘Number of Handguns‘ <chr>, ‘Total Number of Guns‘ <chr>, ...
```

We’re going to specifically look at mass shootings done at a school:

```
school_shooting <- mass_shootings %>% filter(`School Related`=="Yes")
head(school_shooting)
```

```
## # A tibble: 6 x 48
##   CaseID Title       Locat~1 City  State Latit~2 Longi~3 Numbe~4 Total~5 Numbe~6
##    <dbl> <chr>       <chr>   <chr> <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1      1 University~ Austin~ Aust~ Texas    30.2   -97.8      16      17      32
## 2      2 Rose-Mar C~ Mesa, ~ Mesa  Ariz~    33.4  -112.        5       5       1
## 3      4 Clara Bart~ Chicag~ Chic~ Illi~    41.8   -87.7       1       1       3
## 4      5 Olean High~ Olean,~ Olean New ~    42.1   -78.4       3       3       7
## 5      6 Los Angele~ Los An~ Los ~ Cali~    34.2  -119.        1       1       7
## 6      7 Cal State ~ Fuller~ Full~ Cali~    33.9  -118.        7       7       2
## # ... with 38 more variables: ‘Total Number of Victims‘ <dbl>,
## #   Description <chr>, Date <chr>, ‘Day of Week‘ <chr>,
## #   ‘Date - Detailed‘ <chr>, ‘Shooter Name‘ <chr>, ‘Shooter Age(s)‘ <chr>,
## #   ‘Average Shooter Age‘ <chr>, ‘Shooter Sex‘ <chr>, ‘Shooter Race‘ <chr>,
## #   ‘Type of Gun - Detailed‘ <chr>, ‘Type of Gun - General‘ <chr>,
## #   ‘Number of Shotguns‘ <chr>, ‘Number of Rifles‘ <chr>,
## #   ‘Number of Handguns‘ <chr>, ‘Total Number of Guns‘ <chr>, ...
```

Let’s look at the columns of the dataframe:

```
names(school_shooting)
```

```
##  [1] "CaseID"
##  [2] "Title"
##  [3] "Location"
##  [4] "City"
##  [5] "State"
##  [6] "Latitude"
##  [7] "Longitude"
##  [8] "Number of Victim Fatalities"
##  [9] "Total Number of Fatalities"
## [10] "Number of Victims Injured"
## [11] "Total Number of Victims"
## [12] "Description"
## [13] "Date"
## [14] "Day of Week"
## [15] "Date - Detailed"
## [16] "Shooter Name"
## [17] "Shooter Age(s)"
## [18] "Average Shooter Age"
## [19] "Shooter Sex"
## [20] "Shooter Race"
## [21] "Type of Gun - Detailed"
## [22] "Type of Gun - General"
## [23] "Number of Shotguns"
## [24] "Number of Rifles"
## [25] "Number of Handguns"
## [26] "Total Number of Guns"
## [27] "Number of Automatic Guns"
## [28] "Number of Semi-Automatic Guns"
## [29] "Fate of Shooter at the scene"
## [30] "Shooter’s Cause of Death"
## [31] "School Related"
## [32] "Place Type"
## [33] "Relationship to Incident Location"
## [34] "Targeted Victim/s - Detailed"
## [35] "Targeted Victim/s - General"
## [36] "Possible Motive - Detailed"
## [37] "Possible Motive - General"
## [38] "History of Mental Illness - Detailed"
## [39] "History of Mental Illness - General"
## [40] "Data Source 1"
## [41] "Data Source 2"
## [42] "Data Source 3"
## [43] "Data Source 4"
## [44] "Data Source 5"
## [45] "Data Source 6"
## [46] "Data Source 7"
## [47] "Military Experience"
## [48] "Class"
```

Create a subset dataframe from the ‘school shooting’ database:

```
school <- school_shooting %>% select(`State`,`Total Number of Fatalities`,
                            `Day of Week`,
                            `Shooter Age(s)`,
                            `Shooter Race`,
                            `Possible Motive - General`,
                            `History of Mental Illness - General`)
head(school)
```

```
## # A tibble: 6 x 7
##   State      Total Number of Fatalitie~1 Day o~2 Shoot~3 Shoot~4 Possi~5 Histo~6
##   <chr>                            <dbl> <chr>   <chr>   <chr>   <chr>   <chr>
## 1 Texas                               17 Monday  25      White ~ Mental~ Yes
## 2 Arizona                              5 Saturd~ 18      White ~ Mental~ Yes
## 3 Illinois                             1 Thursd~ 14      Unknown Expuls~ Yes
## 4 New York                             3 Monday  17      White ~ Mental~ No
## 5 California                           1 Thursd~ 18      White ~ Social~ Unknown
## 6 California                           7 Monday  37      White ~ Mental~ Yes
## # ... with abbreviated variable names 1: ‘Total Number of Fatalities‘,
## #   2: ‘Day of Week‘, 3: ‘Shooter Age(s)‘, 4: ‘Shooter Race‘,
## #   5: ‘Possible Motive - General‘, 6: ‘History of Mental Illness - General‘
```

Which state has had the most school shootings?

```
school %>% count(`State`) %>% arrange(desc(n))
```

```
## # A tibble: 33 x 2
##    State            n
##    <chr>        <int>
##  1 California      10
##  2 Ohio             5
##  3 Arizona          4
##  4 Illinois         4
##  5 Washington       4
##  6 Michigan         3
##  7 Nevada           3
##  8 New York         3
##  9 Oregon           3
## 10 Pennsylvania     3
## # ... with 23 more rows
```

I’d assume that the main reason why California might seem to have a disproportionately large number of mass shootings is because California has a large number of people. Here is a list of the top ten most populated states in the country:

California (Population: 39,613,493) Texas (Population: 29,730,311) Florida (Population: 21,944,577) New York (Population: 19,299,981) Pennsylvania (Population: 12,804,123) Illinois (Population: 12,569,321) Ohio (Population: 11,714,618) Georgia (Population: 10,830,007) North Carolina (Population: 10,701,022) Michigan (Population: 9,992,427)

What race has the majority of school shooters been?

```
school %>% count(`Shooter Race`) %>% arrange(desc(n))
```

```
## # A tibble: 7 x 2
##   `Shooter Race`                           n
##   <chr>                                <int>
## 1 White American or European American     42
## 2 Black American or African American      14
## 3 Asian American                           7
## 4 Some other race                          5
## 5 Native American or Alaska Native         2
## 6 Unknown                                  2
## 7 Two or more races                        1
```

What age is the most likely age for a person to be a school shooter?

```
school %>% count(`Shooter Age(s)`) %>% arrange(desc(n))
```

```
## # A tibble: 36 x 2
##    `Shooter Age(s)`     n
##    <chr>            <int>
##  1 14                   7
##  2 18                   7
##  3 15                   6
##  4 17                   6
##  5 16                   4
##  6 26                   3
##  7 28                   3
##  8 19                   2
##  9 20                   2
## 10 21                   2
## # ... with 26 more rows
```

There is already a regulation of all handgun purchases by the federal government. All handgun purchases require that you be at least 21 years of age per those federal regulations. Rifles, or long guns as they are usually referred to, are a different story. The individual states have gotten involved when it comes to long guns and, as a result, there are variances from state to state. I think the federal government should also regulate long gun purchases and require the legal age to be 21 years old. Why? The second most common shooter age of the school shooters in the research above was 18 year olds. 18 year olds are simply too young to purchase firearms.