# Tidyverse

Heleine Fouda

2023-11-11

## Getting started

This assignment leverages most of the capabilities built in the tidyverse package.

## Importing the data

```
Url <-read_csv( "https://raw.githubusercontent.com/fivethirtyeight/data/master/hate-crimes/hate_crimes.
```

```
## Rows: 51 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): state
## dbl (11): median_household_income, share_unemployed_seasonal, share_populati...
##
## 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.
```

```
spec(Url)
```

```
## cols(
##   state = col_character(),
##   median_household_income = col_double(),
##   share_unemployed_seasonal = col_double(),
##   share_population_in_metro_areas = col_double(),
##   share_population_with_high_school_degree = col_double(),
##   share_non_citizen = col_double(),
##   share_white_poverty = col_double(),
##   gini_index = col_double(),
##   share_non_white = col_double(),
##   share_voters_voted_trump = col_double(),
##   hate_crimes_per_100k_splc = col_double(),
##   avg_hatecrimes_per_100k_fbi = col_double()
## )
```

```
data <- Url
```

## Data preparation

Removing missing values

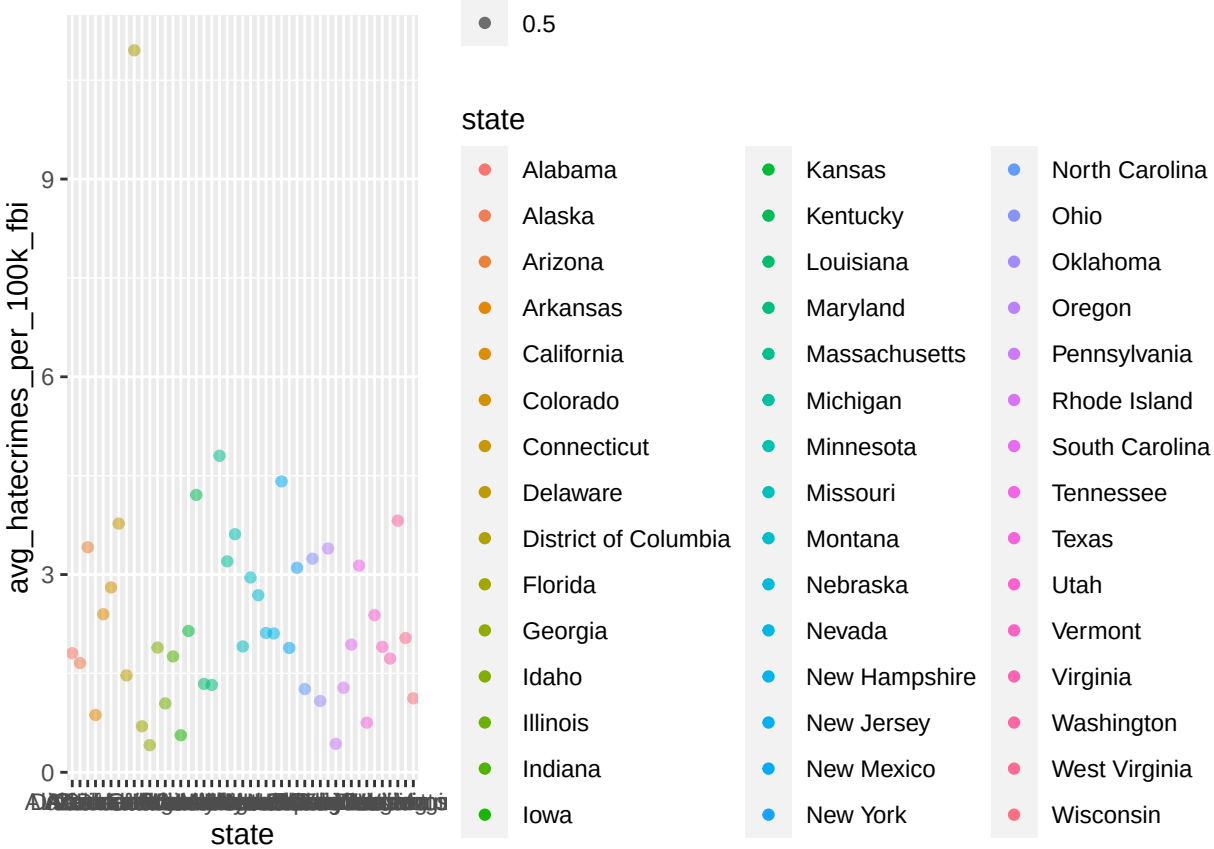
```
hate_crimes <- data[complete.cases(data),]
hate_crimes
```

```
## # A tibble: 45 x 12
##    state    median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
```

```
##    <chr>                <dbl>                <dbl>                <dbl>
##  1 Alabama              42278                0.06                  0.64
##  2 Alaska               67629                0.064                 0.63
##  3 Arizona              49254                0.063                 0.9
##  4 Arkansas             44922                0.052                 0.69
##  5 Califor~             60487                0.059                 0.97
##  6 Colorado             60940                0.04                  0.8
##  7 Connect~             70161                0.052                 0.94
##  8 Delaware             57522                0.049                 0.9
##  9 Distric~             68277                0.067                 1
## 10 Florida              46140                0.052                 0.96
## # i 35 more rows
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

## Data exploration

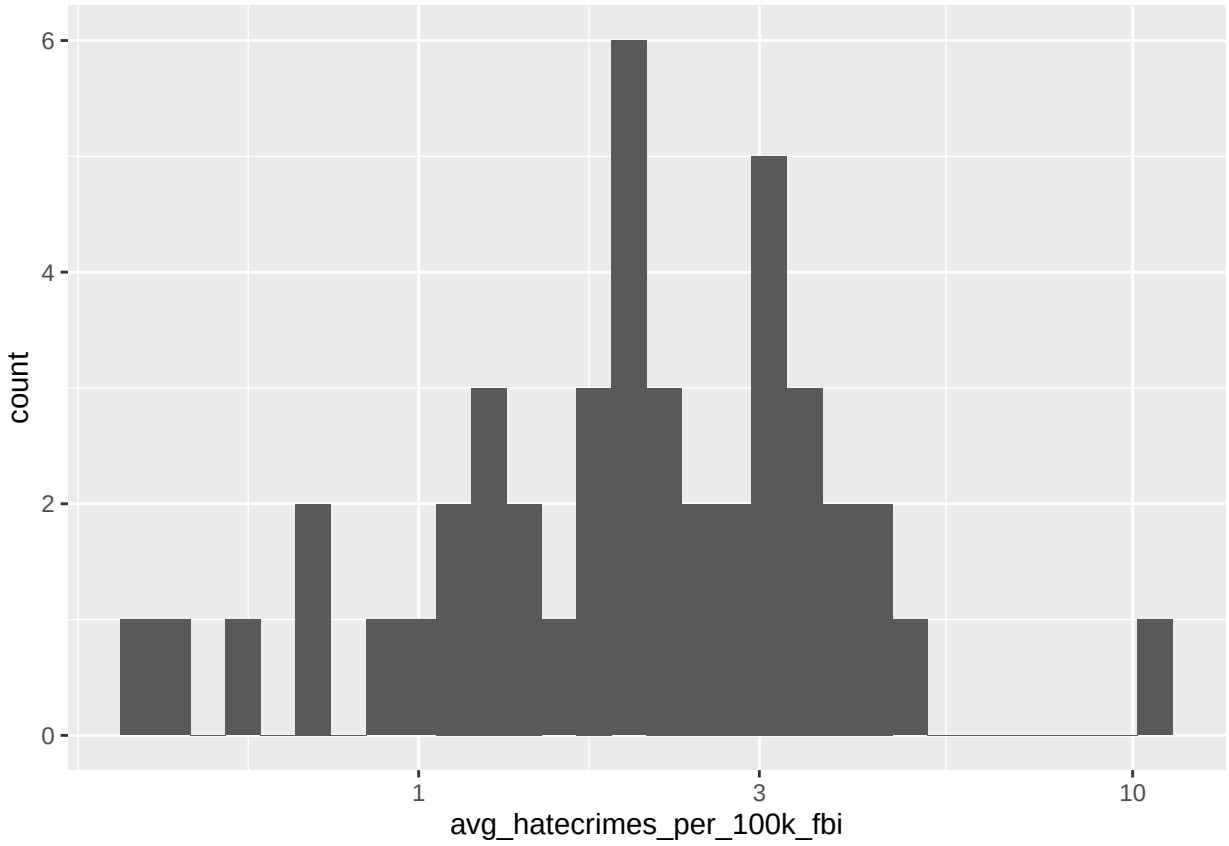
```
# Let's explore the data set with a scatter
ggplot(data = hate_crimes) +
  geom_point(mapping = aes(x = state , y = avg_hatecrimes_per_100k_fbi, color = state, alpha = 0.5))
```



**Let's first get a sense of the hate crimes distribution with a histogram**

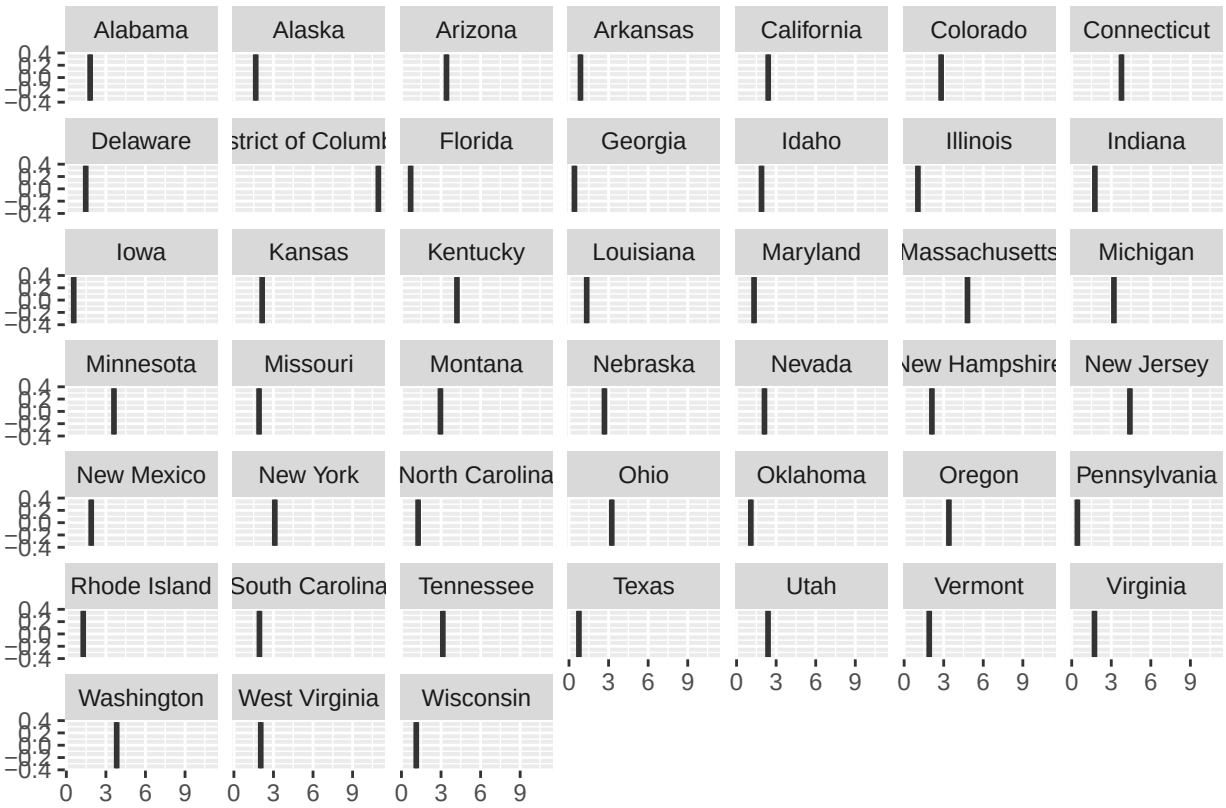
```
ggplot(data = hate_crimes) +
  geom_histogram (mapping = aes(x = avg_hatecrimes_per_100k_fbi)) + scale_x_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



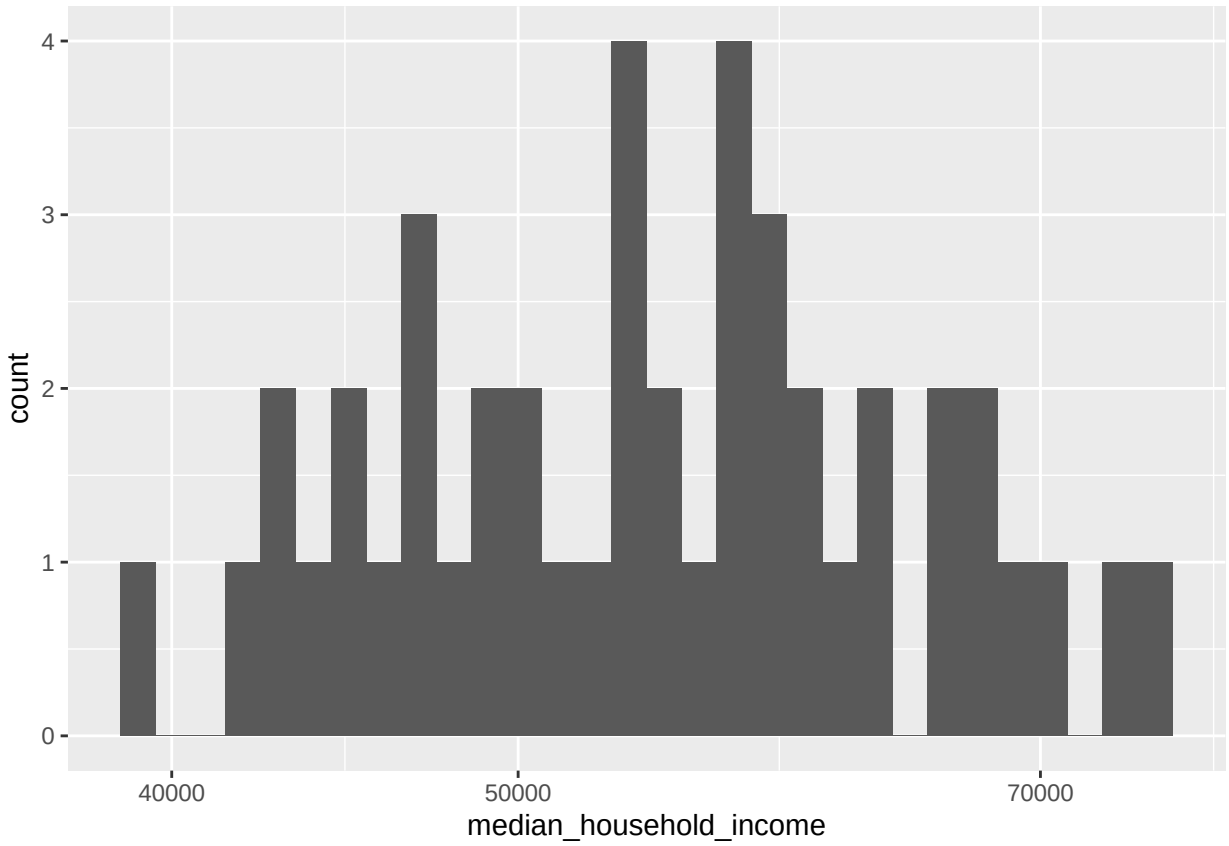

**Now let's get a sense of the hate crimes distribution using a boxplot**

```
ggplot(data = hate_crimes) +
  geom_boxplot (mapping = aes(x = avg_hatecrimes_per_100k_fbi)) +
  facet_wrap(~state)
```




```
# Distribution of household income
ggplot(hate_crimes, aes(x = median_household_income)) +
  geom_histogram()+ scale_x_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




## Descriptive statistics

### Leveraging the deplyr, tydr, and janitor packages

```
glimpse(hate_crimes)
```

```
## Rows: 45
## Columns: 12
## $ state                                 <chr> "Alabama", "Alaska", "Arizona~
## $ median_household_income               <dbl> 42278, 67629, 49254, 44922, 6~
## $ share_unemployed_seasonal             <dbl> 0.060, 0.064, 0.063, 0.052, 0~
## $ share_population_in_metro_areas       <dbl> 0.64, 0.63, 0.90, 0.69, 0.97,~
## $ share_population_with_high_school_degree <dbl> 0.821, 0.914, 0.842, 0.824, 0~
## $ share_non_citizen                     <dbl> 0.02, 0.04, 0.10, 0.04, 0.13,~
## $ share_white_poverty                   <dbl> 0.12, 0.06, 0.09, 0.12, 0.09,~
## $ gini_index                            <dbl> 0.472, 0.422, 0.455, 0.458, 0~
## $ share_non_white                       <dbl> 0.35, 0.42, 0.49, 0.26, 0.61,~
## $ share_voters_voted_trump              <dbl> 0.63, 0.53, 0.50, 0.60, 0.33,~
## $ hate_crimes_per_100k_splc             <dbl> 0.12583893, 0.14374012, 0.225~
## $ avg_hatecrimes_per_100k_fbi           <dbl> 1.8064105, 1.6567001, 3.41392~
```

```
dim(hate_crimes)
```

```
## [1] 45 12
```

```
str(hate_crimes)
```

```
## tibble [45 x 12] (S3: tbl_df/tbl/data.frame)
```

```
##  $ state                                 : chr [1:45] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ median_household_income               : num [1:45] 42278 67629 49254 44922 60487 ...
##  $ share_unemployed_seasonal             : num [1:45] 0.06 0.064 0.063 0.052 0.059 0.04 0.052 0.049
##  $ share_population_in_metro_areas       : num [1:45] 0.64 0.63 0.9 0.69 0.97 0.8 0.94 0.9 1 0.96
##  $ share_population_with_high_school_degree: num [1:45] 0.821 0.914 0.842 0.824 0.806 0.893 0.886 0.8
##  $ share_non_citizen                     : num [1:45] 0.02 0.04 0.1 0.04 0.13 0.06 0.06 0.05 0.11 0
##  $ share_white_poverty                   : num [1:45] 0.12 0.06 0.09 0.12 0.09 0.07 0.06 0.08 0.04
##  $ gini_index                            : num [1:45] 0.472 0.422 0.455 0.458 0.471 0.457 0.486 0.4
##  $ share_non_white                       : num [1:45] 0.35 0.42 0.49 0.26 0.61 0.31 0.3 0.37 0.63 0
##  $ share_voters_voted_trump              : num [1:45] 0.63 0.53 0.5 0.6 0.33 0.44 0.41 0.42 0.04 0
##  $ hate_crimes_per_100k_splc             : num [1:45] 0.1258 0.1437 0.2253 0.0691 0.2558 ...
##  $ avg_hatecrimes_per_100k_fbi           : num [1:45] 1.806 1.657 3.414 0.869 2.398 ...
```

```
names(hate_crimes)
```

```
##  [1] "state"
##  [2] "median_household_income"
##  [3] "share_unemployed_seasonal"
##  [4] "share_population_in_metro_areas"
##  [5] "share_population_with_high_school_degree"
##  [6] "share_non_citizen"
##  [7] "share_white_poverty"
##  [8] "gini_index"
##  [9] "share_non_white"
## [10] "share_voters_voted_trump"
## [11] "hate_crimes_per_100k_splc"
## [12] "avg_hatecrimes_per_100k_fbi"
```

```
# We can also quickly add our tally values to our tibble using add_tally().

 hate_crimes %>%
  add_tally() %>%
  glimpse()
```

```
## Rows: 45
## Columns: 13
## $ state                                 <chr> "Alabama", "Alaska", "Arizona~
## $ median_household_income               <dbl> 42278, 67629, 49254, 44922, 6~
## $ share_unemployed_seasonal             <dbl> 0.060, 0.064, 0.063, 0.052, 0~
## $ share_population_in_metro_areas       <dbl> 0.64, 0.63, 0.90, 0.69, 0.97,~
## $ share_population_with_high_school_degree <dbl> 0.821, 0.914, 0.842, 0.824, 0~
## $ share_non_citizen                     <dbl> 0.02, 0.04, 0.10, 0.04, 0.13,~
## $ share_white_poverty                   <dbl> 0.12, 0.06, 0.09, 0.12, 0.09,~
## $ gini_index                            <dbl> 0.472, 0.422, 0.455, 0.458, 0~
## $ share_non_white                       <dbl> 0.35, 0.42, 0.49, 0.26, 0.61,~
## $ share_voters_voted_trump              <dbl> 0.63, 0.53, 0.50, 0.60, 0.33,~
## $ hate_crimes_per_100k_splc             <dbl> 0.12583893, 0.14374012, 0.225~
## $ avg_hatecrimes_per_100k_fbi           <dbl> 1.8064105, 1.6567001, 3.41392~
## $ n                                     <int> 45, 45, 45, 45, 45, 45, 45, 4~
```

**Getting a quick summary of the data frame using the skimr package.**

```
skim(hate_crimes)
```

Table 1: Data summary

| | |
|---|---|
| Name | hate_crimes |
| Number of rows | 45 |
| Number of columns | 12 |
| Column type frequency: | |
| character | 1 |
| numeric | 11 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| state | 0 | 1 | 4 | 20 | 0 | 45 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| median_household_income | 0 | 1 | 55299.49 | 8979.49 | 39552.00 | 48060.00 | 54916.00 | 60708.00 | 76165.00 | |
| share_unemployed_seasonal | 0 | 1 | 0.05 | 0.01 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | |
| share_population_in_metro_areas | 0 | 1 | 0.78 | 0.16 | 0.34 | 0.69 | 0.81 | 0.90 | 1.00 | |
| share_population_with_high_school_degree | 0 | 1 | 0.87 | 0.03 | 0.80 | 0.84 | 0.87 | 0.89 | 0.92 | |
| share_non_citizen | 0 | 1 | 0.06 | 0.03 | 0.01 | 0.03 | 0.05 | 0.08 | 0.13 | |
| share_white_poverty | 0 | 1 | 0.09 | 0.02 | 0.04 | 0.07 | 0.09 | 0.10 | 0.17 | |
| gini_index | 0 | 1 | 0.46 | 0.02 | 0.42 | 0.44 | 0.46 | 0.47 | 0.53 | |
| share_non_white | 0 | 1 | 0.32 | 0.15 | 0.06 | 0.21 | 0.30 | 0.42 | 0.63 | |
| share_voters_voted_trump | 0 | 1 | 0.48 | 0.11 | 0.04 | 0.41 | 0.49 | 0.57 | 0.69 | |
| hate_crimes_per_100k_splc | 0 | 1 | 0.30 | 0.25 | 0.07 | 0.14 | 0.23 | 0.35 | 1.52 | |
| avg_hatecrimes_per_100k_fbi | 0 | 1 | 2.37 | 1.72 | 0.41 | 1.32 | 1.94 | 3.14 | 10.95 | |

```
## # see summary for specified columns
skim(msleep, genus, vore, sleep_total)
```

Table 4: Data summary

| | |
|---|---|
| Name | msleep |
| Number of rows | 83 |
| Number of columns | 11 |
| Column type frequency: | |
| character | 2 |
| numeric | 1 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| genus | 0 | 1.00 | 3 | 13 | 0 | 77 | 0 |
| vore | 7 | 0.92 | 4 | 7 | 0 | 4 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| sleep_total | 0 | 1 | 10.43 | 4.45 | 1.9 | 7.85 | 10.1 | 13.75 | 19.9 | |

```
# summarizing across
sum1 <- hate_crimes %>%
  summarize(across(state:avg_hatecrimes_per_100k_fbi, mean, na.rm = TRUE))
```

```
## Warning: There were 2 warnings in `summarize()`.
## The first warning was:
## i In argument: `across(state:avg_hatecrimes_per_100k_fbi, mean, na.rm = TRUE)`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
## i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```
sum1
```

```
## # A tibble: 1 x 12
##   state median_household_income share_unemployed_seasonal share_population_in_~1
##   <dbl>                  <dbl>                      <dbl>                   <dbl>
## 1    NA                 55299.                     0.0508                   0.782
## # i abbreviated name: 1: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

```
summary(hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.412   1.325   1.937   2.374   3.136  10.953
```

```
# Summarizing the average hate crimes using the janitor package and the function tabyl
summ_crimes <- hate_crimes %>%
  tabyl(avg_hatecrimes_per_100k_fbi)
summ_crimes
```

```
##  avg_hatecrimes_per_100k_fbi n    percent
##                    0.4120118 1 0.02222222
##                    0.4309276 1 0.02222222
##                    0.5613956 1 0.02222222
##                    0.6980703 1 0.02222222
```

```
##                    0.7527683 1 0.02222222
##                    0.8692089 1 0.02222222
##                    1.0440158 1 0.02222222
##                    1.0816721 1 0.02222222
##                    1.1219447 1 0.02222222
##                    1.2626798 1 0.02222222
##                    1.2825718 1 0.02222222
##                    1.3248395 1 0.02222222
##                    1.3411696 1 0.02222222
##                    1.4699796 1 0.02222222
##                    1.6567001 1 0.02222222
##                    1.7247546 1 0.02222222
##                    1.7573566 1 0.02222222
##                    1.8064105 1 0.02222222
##                    1.8864352 1 0.02222222
##                    1.8913305 1 0.02222222
##                    1.9030814 1 0.02222222
##                    1.9089550 1 0.02222222
##                    1.9370828 1 0.02222222
##                    2.0370536 1 0.02222222
##                    2.1059886 1 0.02222222
##                    2.1139902 1 0.02222222
##                    2.1439867 1 0.02222222
##                    2.3840650 1 0.02222222
##                    2.3979859 1 0.02222222
##                    2.6862484 1 0.02222222
##                    2.8046888 1 0.02222222
##                    2.9549594 1 0.02222222
##                    3.1021643 1 0.02222222
##                    3.1360512 1 0.02222222
##                    3.2004423 1 0.02222222
##                    3.2404204 1 0.02222222
##                    3.3948861 1 0.02222222
##                    3.4139280 1 0.02222222
##                    3.6124118 1 0.02222222
##                    3.7727015 1 0.02222222
##                    3.8177403 1 0.02222222
##                    4.2078896 1 0.02222222
##                    4.4132026 1 0.02222222
##                    4.8018993 1 0.02222222
##                   10.9534797 1 0.02222222
```

```
# Note, that tabyl assumes categorical variables.
```

```
summary(hate_crimes$share_non_white)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0600  0.2100  0.3000  0.3176  0.4200  0.6300
```

### Let's do some filtering using dplyr

```
# Let's filter data set rows to only include  households with a median income equal or less than 50000

low_income_states <-hate_crimes %>%
  dplyr::filter (median_household_income <= 45000)
```

```
low_income_states
```

```
## # A tibble: 7 x 12
##   state     median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 Alabama                    42278                  0.06                    0.64
## 2 Arkansas                   44922                  0.052                   0.69
## 3 Kentucky                   42786                  0.05                    0.56
## 4 Louisiana                  42406                  0.06                    0.81
## 5 South Ca~                  44929                  0.057                   0.79
## 6 Tennessee                  43716                  0.057                   0.82
## 7 West Vir~                  39552                  0.073                   0.55
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

```
low_income_states %>%
  arrange(desc(median_household_income))
```

```
## # A tibble: 7 x 12
##   state     median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 South Ca~                  44929                  0.057                   0.79
## 2 Arkansas                   44922                  0.052                   0.69
## 3 Tennessee                  43716                  0.057                   0.82
## 4 Kentucky                   42786                  0.05                    0.56
## 5 Louisiana                  42406                  0.06                    0.81
## 6 Alabama                    42278                  0.06                    0.64
## 7 West Vir~                  39552                  0.073                   0.55
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

```
low_income_states
```

```
## # A tibble: 7 x 12
##   state     median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 Alabama                    42278                  0.06                    0.64
## 2 Arkansas                   44922                  0.052                   0.69
## 3 Kentucky                   42786                  0.05                    0.56
## 4 Louisiana                  42406                  0.06                    0.81
## 5 South Ca~                  44929                  0.057                   0.79
## 6 Tennessee                  43716                  0.057                   0.82
## 7 West Vir~                  39552                  0.073                   0.55
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
```

```
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

```
# The 7 states with the lowest household income
  (slice_tail(low_income_states , n=7))
```

```
## # A tibble: 7 x 12
##   state     median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 Alabama                    42278                  0.06                    0.64
## 2 Arkansas                   44922                  0.052                   0.69
## 3 Kentucky                   42786                  0.05                    0.56
## 4 Louisiana                  42406                  0.06                    0.81
## 5 South Ca~                  44929                  0.057                   0.79
## 6 Tennessee                  43716                  0.057                   0.82
## 7 West Vir~                  39552                  0.073                   0.55
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

**Let's filter both specific observations or rows and specific cases or columns that are of interest to us**

```
# Filter data set columns to only include ...
low_income_df <- low_income_states %>% select(state,avg_hatecrimes_per_100k_fbi, median_household_income
low_income_df
```

```
## # A tibble: 7 x 8
##   state     avg_hatecrimes_per_1~1 median_household_inc~2 share_population_wit~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 Alabama                    1.81                   42278                  0.821
## 2 Arkansas                   0.869                  44922                  0.824
## 3 Kentucky                   4.21                   42786                  0.817
## 4 Louisiana                  1.34                   42406                  0.822
## 5 South Ca~                  1.94                   44929                  0.836
## 6 Tennessee                  3.14                   43716                  0.831
## 7 West Vir~                  2.04                   39552                  0.828
## # i abbreviated names: 1: avg_hatecrimes_per_100k_fbi,
## #   2: median_household_income, 3: share_population_with_high_school_degree
## # i 4 more variables: share_white_poverty <dbl>, share_non_white <dbl>,
## #   share_voters_voted_trump <dbl>, share_non_citizen <dbl>
```

```
 # Let's filter data set rows to only include  households with a median income equal or superior to 650

middle_income_states <-hate_crimes %>%
dplyr::filter(median_household_income >= 66000)
```

```
# The 7 states with the highest household income
  (slice_head(middle_income_states , n=7))
```

```
## # A tibble: 7 x 12
##   state     median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
```

```
## 1 Alaska                    67629                  0.064                  0.63
## 2 Connecti~                 70161                  0.052                  0.94
## 3 District~                 68277                  0.067                  1
## 4 Maryland                  76165                  0.051                  0.97
## 5 Minnesota                 67244                  0.038                  0.75
## 6 New Hamp~                 73397                  0.034                  0.63
## 7 Virginia                  66155                  0.043                  0.89
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

```
# alternatively
arrange(middle_income_states)
```

```
## # A tibble: 7 x 12
##   state     median_household_inc~1 share_unemployed_sea~2 share_population_in_~3
##   <chr>                      <dbl>                  <dbl>                   <dbl>
## 1 Alaska                     67629                  0.064                    0.63
## 2 Connecti~                  70161                  0.052                    0.94
## 3 District~                  68277                  0.067                    1
## 4 Maryland                   76165                  0.051                    0.97
## 5 Minnesota                  67244                  0.038                    0.75
## 6 New Hamp~                  73397                  0.034                    0.63
## 7 Virginia                   66155                  0.043                    0.89
## # i abbreviated names: 1: median_household_income,
## #   2: share_unemployed_seasonal, 3: share_population_in_metro_areas
## # i 8 more variables: share_population_with_high_school_degree <dbl>,
## #   share_non_citizen <dbl>, share_white_poverty <dbl>, gini_index <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   hate_crimes_per_100k_splc <dbl>, avg_hatecrimes_per_100k_fbi <dbl>
```

Now let's take the original data set and then group it by states before re-arranging it in descending order based on their average crime rates

```
# Grouping the data set by state and re-arranging them in descending order based on their average hate
 states <- hate_crimes %>%
  group_by(state) %>%
  select(state, avg_hatecrimes_per_100k_fbi) %>%
  summarize(N=n(), mean_hatecrimes = avg_hatecrimes_per_100k_fbi) %>%
  arrange(desc(mean_hatecrimes))
states
```

```
## # A tibble: 45 x 3
##    state                   N mean_hatecrimes
##    <chr>               <int>           <dbl>
##  1 District of Columbia     1           11.0
##  2 Massachusetts            1            4.80
##  3 New Jersey               1            4.41
##  4 Kentucky                 1            4.21
##  5 Washington               1            3.82
##  6 Connecticut              1            3.77
##  7 Minnesota                1            3.61
##  8 Arizona                  1            3.41
```

```
##  9 Oregon                   1            3.39
## 10 Ohio                     1            3.24
## # i 35 more rows
```

**Let's select the 7 states with the lowest average hate crime rates.**

```
lowest_hcrimes<- states %>%
  slice_tail(n = 7)
 lowest_hcrimes
```

```
## # A tibble: 7 x 3
##   state            N mean_hatecrimes
##   <chr>        <int>           <dbl>
## 1 Illinois         1           1.04
## 2 Arkansas         1           0.869
## 3 Texas            1           0.753
## 4 Florida          1           0.698
## 5 Iowa             1           0.561
## 6 Pennsylvania     1           0.431
## 7 Georgia          1           0.412
```

```
sum_lowest_hcrimes <- lowest_hcrimes %>%
summarize(across(state:mean_hatecrimes, mean, na.rm = TRUE))
```

```
## Warning: There was 1 warning in `summarize()`.
## i In argument: `across(state:mean_hatecrimes, mean, na.rm = TRUE)`.
## Caused by warning in `mean.default()`:
## ! argument is not numeric or logical: returning NA
```

```
sum_lowest_hcrimes
```

```
## # A tibble: 1 x 3
##   state     N mean_hatecrimes
##   <dbl> <dbl>           <dbl>
## 1    NA     1           0.681
```

**Let's select the 7 states with the highest average hate crime rates, using dplyr**

```
highest_hcrimes<- states %>%
  slice_head(n = 7)
 highest_hcrimes
```

```
## # A tibble: 7 x 3
##   state                   N mean_hatecrimes
##   <chr>               <int>           <dbl>
## 1 District of Columbia     1           11.0
## 2 Massachusetts            1            4.80
## 3 New Jersey               1            4.41
## 4 Kentucky                 1            4.21
## 5 Washington               1            3.82
## 6 Connecticut              1            3.77
## 7 Minnesota                1            3.61
```

```
sum_highest_crimes <- highest_hcrimes %>%
summarize(across(state:mean_hatecrimes, mean, na.rm = TRUE))
```

```
## Warning: There was 1 warning in `summarize()`.
```

```
## i In argument: `across(state:mean_hatecrimes, mean, na.rm = TRUE)`.
## Caused by warning in `mean.default()`:
## ! argument is not numeric or logical: returning NA
```

```
sum_highest_crimes
```

```
## # A tibble: 1 x 3
##   state     N mean_hatecrimes
##   <dbl> <dbl>           <dbl>
## 1    NA     1            5.08
```

### Some states with low household income score high in hate crimes

```
Low_income_states <-low_income_df%>%
  dplyr::filter(avg_hatecrimes_per_100k_fbi >= 20000536)
view(Low_income_states)
```

Low income states with the highest average rate crime:

```
Low_income1 <-low_income_df%>%
  dplyr::filter(avg_hatecrimes_per_100k_fbi >= 20000536) %>%
  arrange(desc(avg_hatecrimes_per_100k_fbi))
Low_income1
```

```
## # A tibble: 0 x 8
## # i 8 variables: state <chr>, avg_hatecrimes_per_100k_fbi <dbl>,
## #   median_household_income <dbl>,
## #   share_population_with_high_school_degree <dbl>, share_white_poverty <dbl>,
## #   share_non_white <dbl>, share_voters_voted_trump <dbl>,
## #   share_non_citizen <dbl>
```

### Let's explore the relationship between average hate crimes and the other variables within the data set. Which factor best explain hate crimes rate?

Controlling for education

```
# Controlling for secondary education
High_school <- hate_crimes %>%
  group_by(state) %>%
  select(share_population_with_high_school_degree, avg_hatecrimes_per_100k_fbi) %>%
              arrange(desc(share_population_with_high_school_degree))
```

```
## Adding missing grouping variables: `state`
```

```
 High_school
```

```
## # A tibble: 45 x 3
## # Groups:   state [45]
##    state         share_population_with_high_school_degree avg_hatecrimes_per_1~1
##    <chr>                                            <dbl>                  <dbl>
##  1 Minnesota                                        0.915                  3.61
##  2 Alaska                                           0.914                  1.66
##  3 Iowa                                             0.914                  0.561
##  4 New Hampshire                                    0.913                  2.11
##  5 Vermont                                          0.91                   1.90
##  6 Montana                                          0.908                  2.95
##  7 Utah                                             0.904                  2.38
```

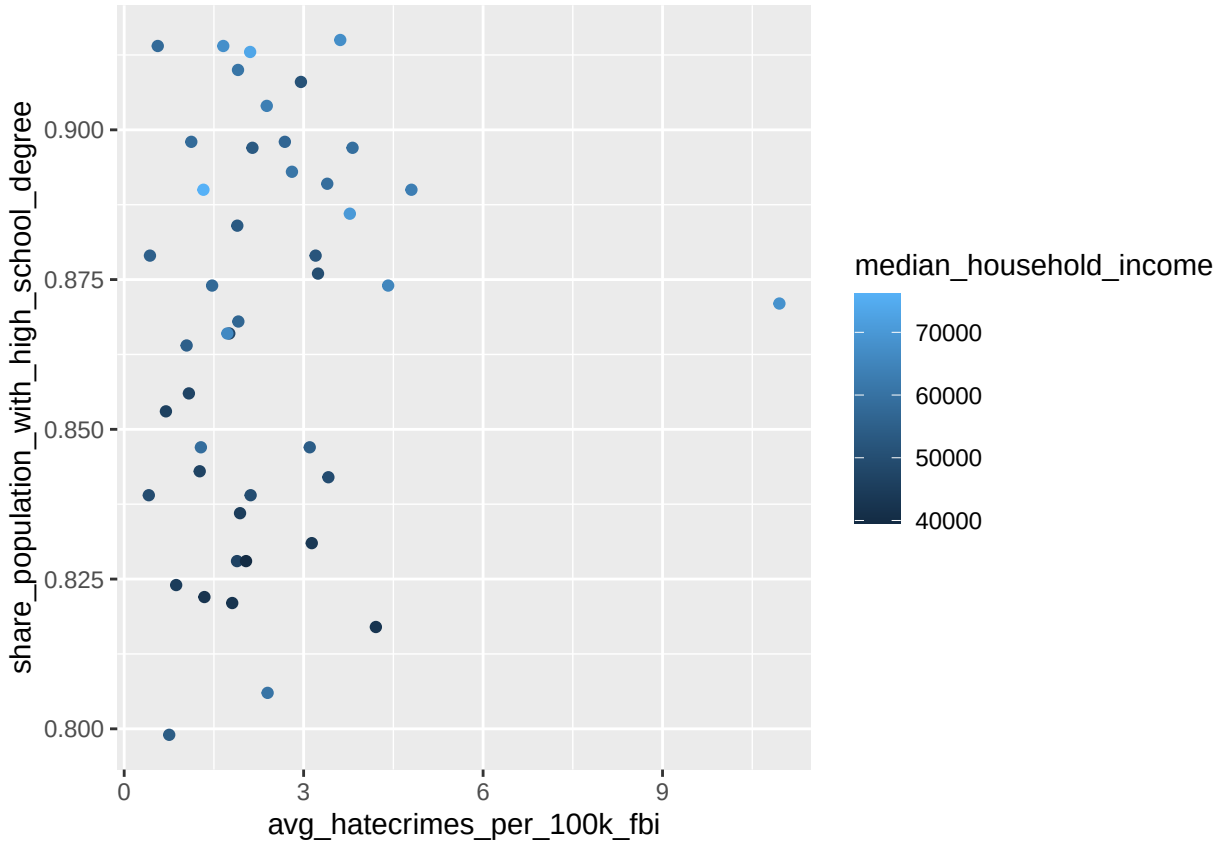
```
##  8 Nebraska                                    0.898                 2.69
##  9 Wisconsin                                   0.898                 1.12
## 10 Kansas                                      0.897                 2.14
## # i 35 more rows
## # i abbreviated name: 1: avg_hatecrimes_per_100k_fbi
```

There is a positive but very weak relationship between hate crime rates and secondary education

```
cor(hate_crimes$share_population_with_high_school_degree, hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] 0.1405676
```

```
# Controlling for secondary education
ggplot(hate_crimes, aes(x = avg_hatecrimes_per_100k_fbi, y = share_population_with_high_school_degree,
  geom_point()
```




Controlling for the share of non white population: very weak positive relationship

```
# controlling for the share of non white population
 non_white <- hate_crimes %>%
  group_by(state) %>%
  select(share_non_white, avg_hatecrimes_per_100k_fbi) %>%
              arrange(desc(share_non_white))
```

```
## Adding missing grouping variables: `state`
```

```
   non_white
```

```
## # A tibble: 45 x 3
## # Groups:   state [45]
##    state                share_non_white avg_hatecrimes_per_100k_fbi
```

```
##    <chr>                          <dbl>                      <dbl>
##  1 District of Columbia            0.63                     11.0
##  2 New Mexico                      0.62                      1.89
##  3 California                      0.61                      2.40
##  4 Texas                           0.56                      0.753
##  5 Maryland                        0.5                       1.32
##  6 Nevada                          0.5                       2.11
##  7 Arizona                         0.49                      3.41
##  8 Georgia                         0.48                      0.412
##  9 Florida                         0.46                      0.698
## 10 New Jersey                      0.44                      4.41
## # i 35 more rows
```

```
cor(hate_crimes$share_non_white, hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] 0.1345048
```

Controlling for share of non citizen: Weak positive relationship here but there seems to be an outlier represented by the District of Columbia

```
# Controlling for share of non citizen
non_citizen <- hate_crimes %>%
  group_by(state) %>%
  select(share_voters_voted_trump, share_white_poverty, share_non_citizen , avg_hatecrimes_per_100k_fbi
              arrange(desc(share_non_citizen ))
```

```
## Adding missing grouping variables: `state`
```

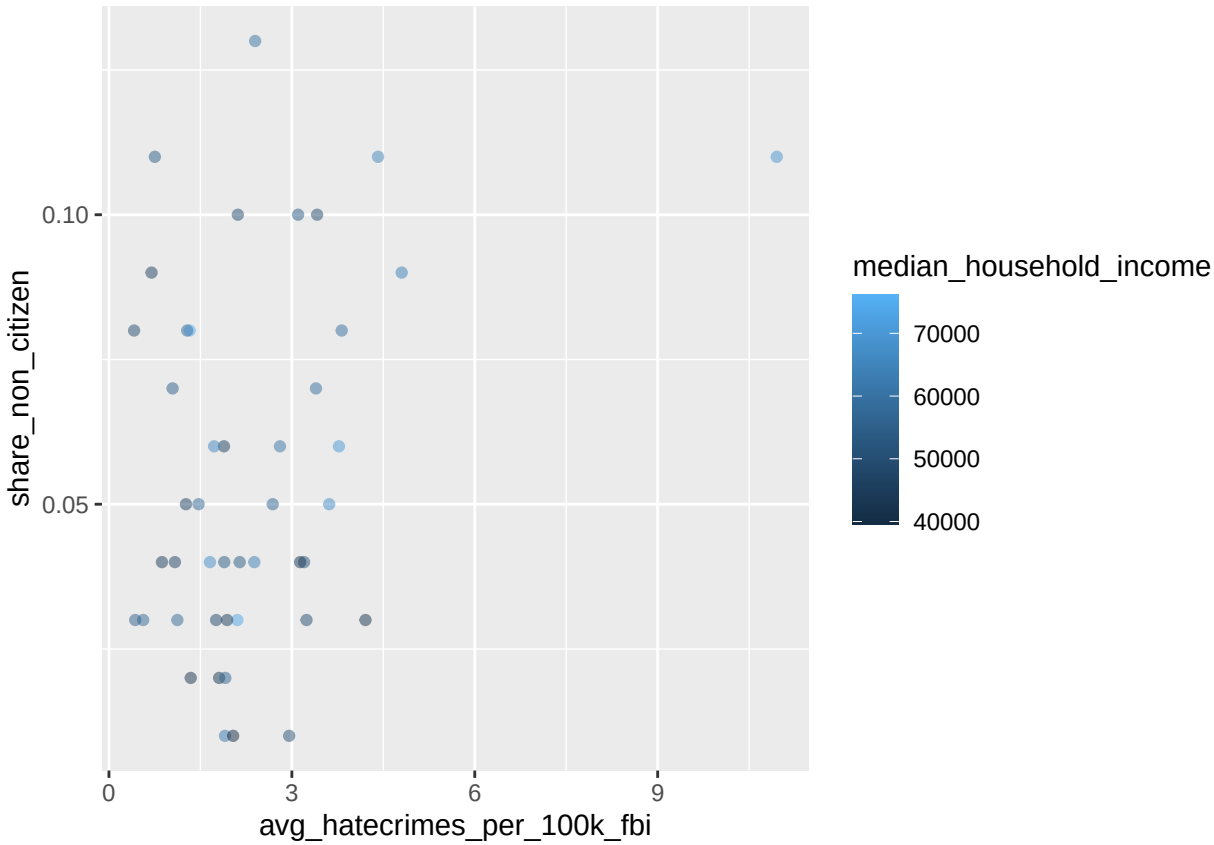
```
   non_citizen
```

```
## # A tibble: 45 x 5
## # Groups:   state [45]
##    state            share_voters_voted_t~1 share_white_poverty share_non_citizen
##    <chr>                             <dbl>               <dbl>             <dbl>
##  1 California                         0.33                0.09              0.13
##  2 District of Col~                   0.04                0.04              0.11
##  3 New Jersey                         0.42                0.07              0.11
##  4 Texas                              0.53                0.08              0.11
##  5 Arizona                            0.5                 0.09              0.1
##  6 Nevada                             0.46                0.08              0.1
##  7 New York                           0.37                0.1               0.1
##  8 Florida                            0.49                0.11              0.09
##  9 Massachusetts                      0.34                0.08              0.09
## 10 Georgia                            0.51                0.09              0.08
## # i 35 more rows
## # i abbreviated name: 1: share_voters_voted_trump
## # i 1 more variable: avg_hatecrimes_per_100k_fbi <dbl>
```

```
cor(hate_crimes$share_non_citizen , hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] 0.3125537
```

```
ggplot(hate_crimes, aes(x= avg_hatecrimes_per_100k_fbi, y = share_non_citizen, color= median_household_
  geom_point(alpha = 0.5)
```

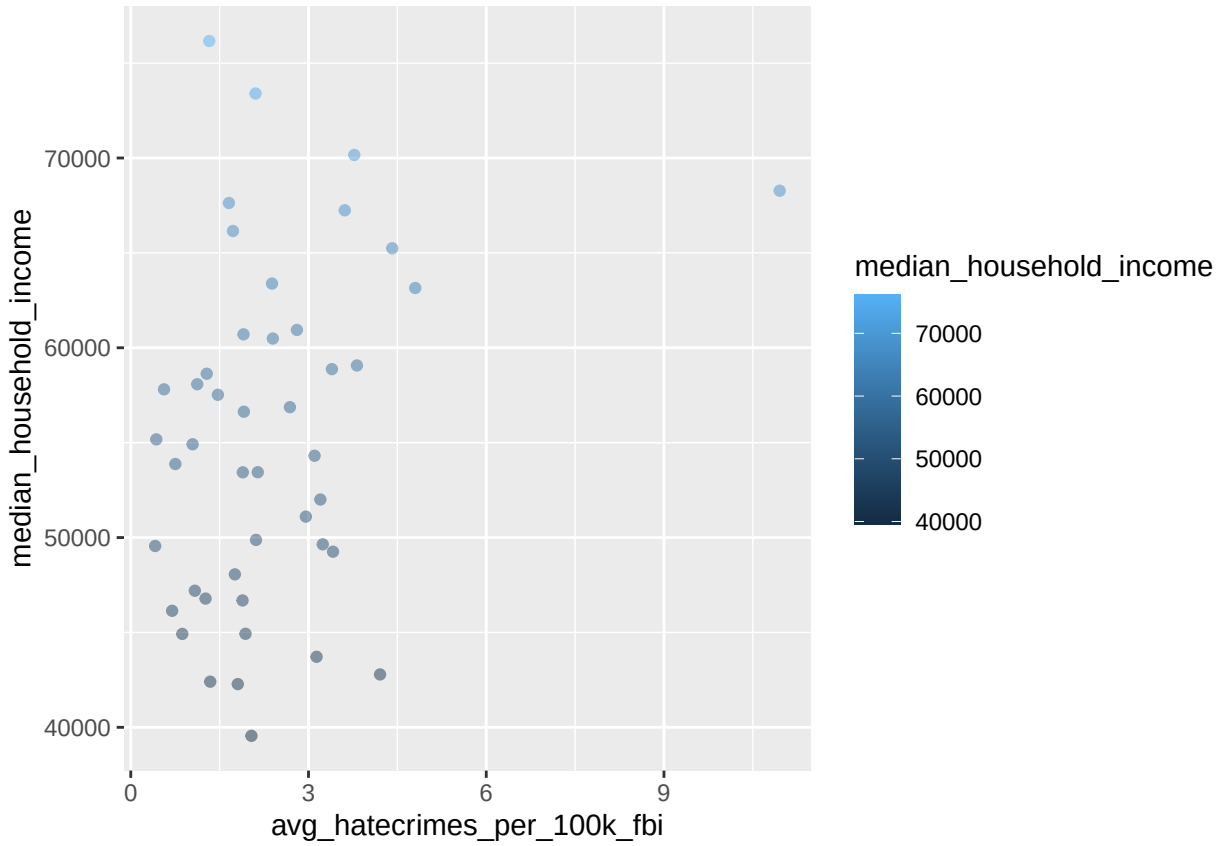



controlling for household income: a weak but positive relationship

```
# controlling for household income
cor(hate_crimes$median_household_income , hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] 0.2906101
```

```
# controlling for household income
ggplot(hate_crimes, aes(x = avg_hatecrimes_per_100k_fbi,  y = median_household_income, color = median_h
  geom_point(alpha= 0.5)
```




```
facet_wrap(~median_household_income)
```

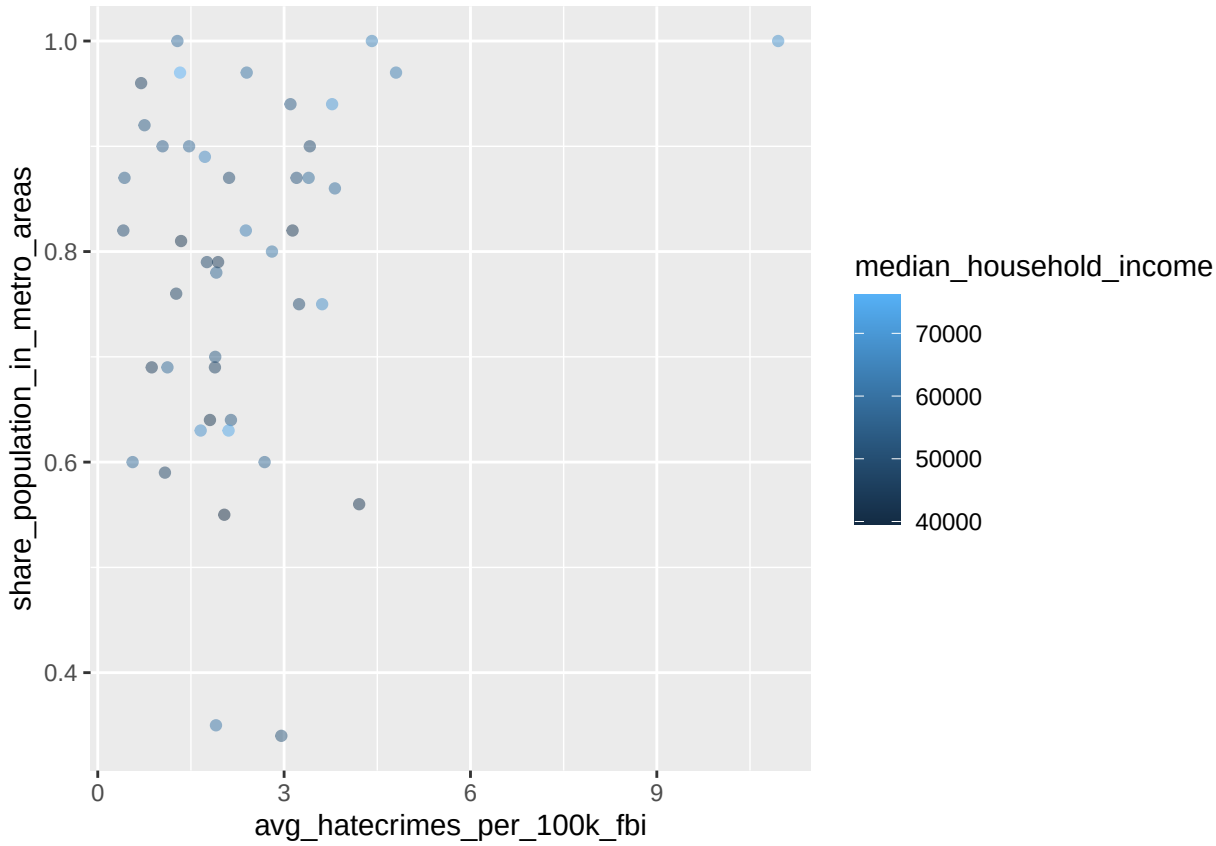
```
## <ggproto object: Class FacetWrap, Facet, gg>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map_data: function
##     params: list
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train_scales: function
##     vars: function
##     super:  <ggproto object: Class FacetWrap, Facet, gg>
```

Controlling for population in metropolitan areas: A weak but positive relationship

```
cor(hate_crimes$share_population_in_metro_areas , hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] 0.21617
```

```
ggplot(hate_crimes, aes(x = avg_hatecrimes_per_100k_fbi, y =  share_population_in_metro_areas,color = m
  geom_point(alpha = 0.5)
```

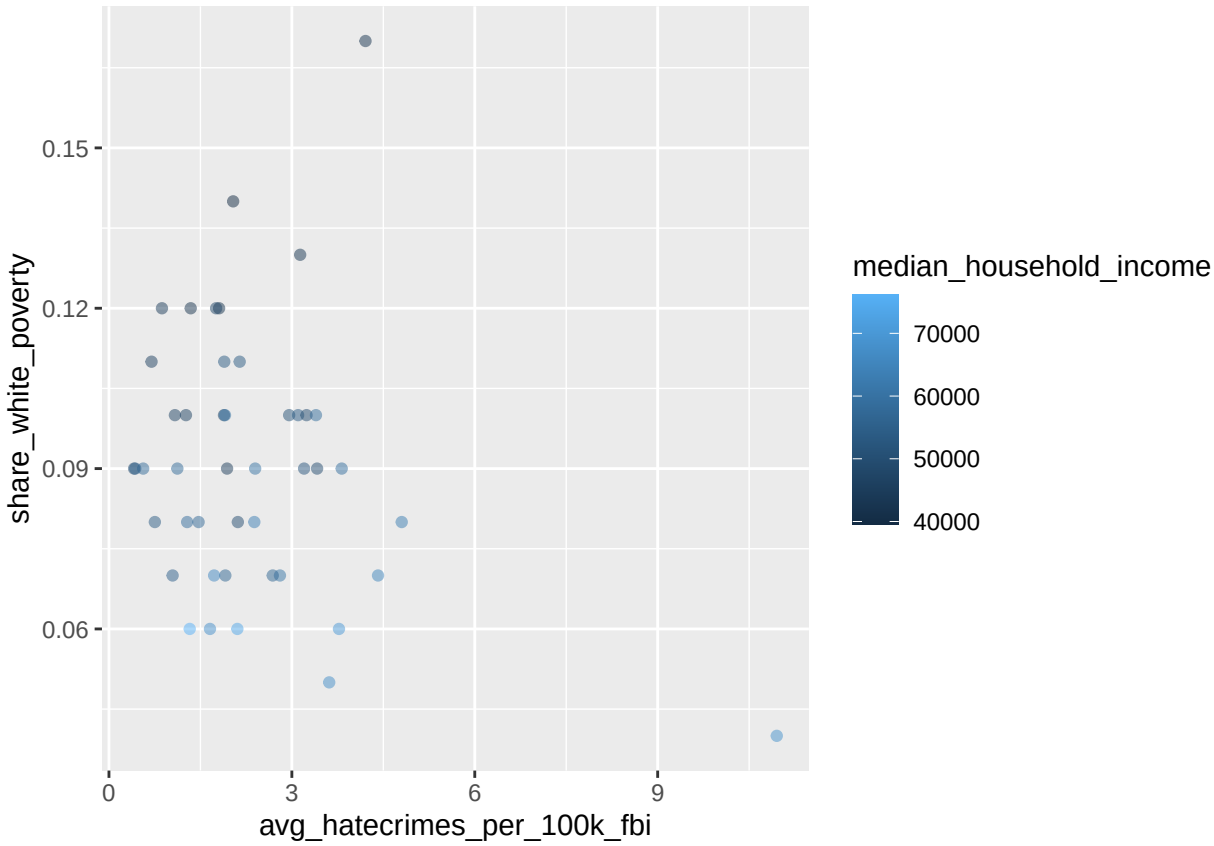



Controlling for share_white_poverty: There is a negative relationship between hate crimes and the level of white poverty

```
cor(hate_crimes$share_white_poverty, hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] -0.2426443
```

```
ggplot(hate_crimes, aes(x = avg_hatecrimes_per_100k_fbi,
y =share_white_poverty, color= median_household_income))+
  geom_point(alpha = 0.5)
```

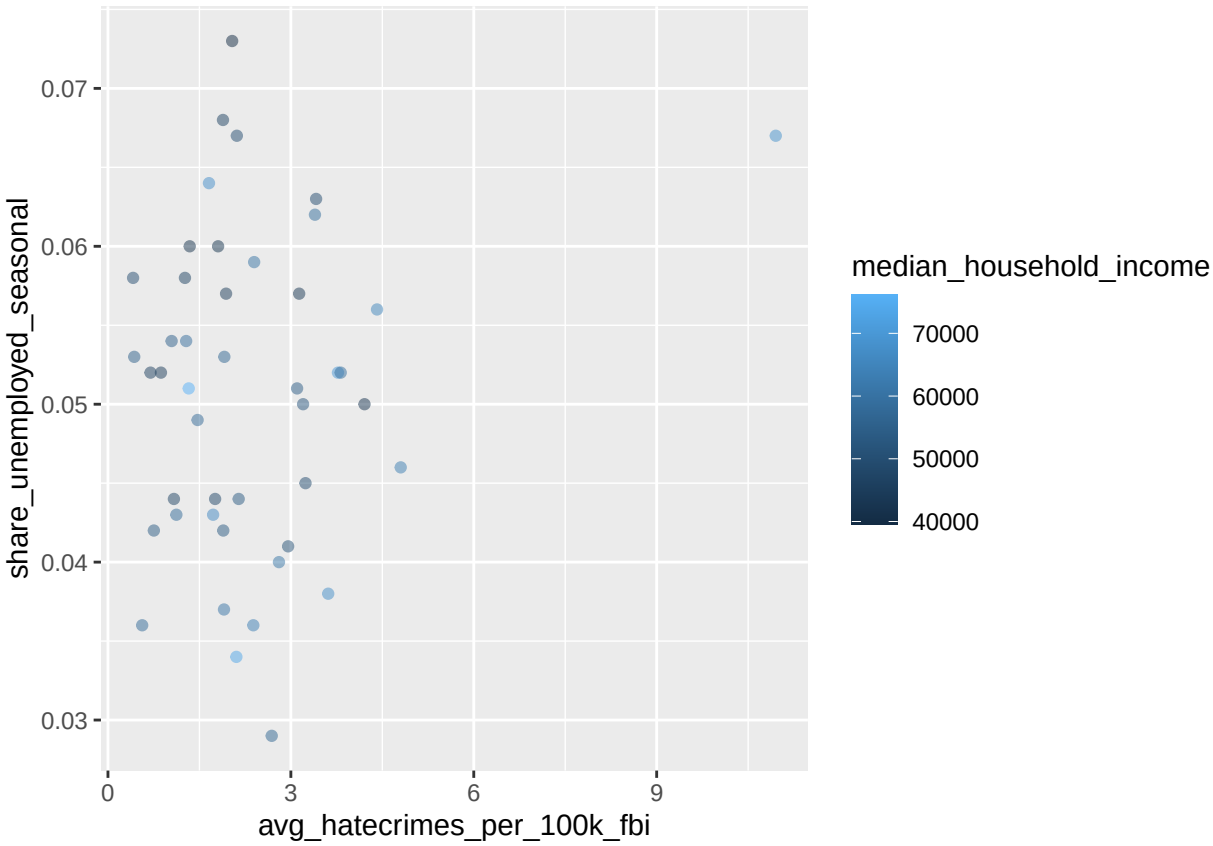



Controlling for seasonal unemployment: A very weak but positive realtionship

```
cor(hate_crimes$share_unemployed_seasonal, hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] 0.1721765
```

```
ggplot(hate_crimes, aes(x = avg_hatecrimes_per_100k_fbi, y= share_unemployed_seasonal, color = median_h
  geom_point(alpha = 0.5)
```




controlling for Trump votes: a strong but negative relationship

```
cor(hate_crimes$share_voters_voted_trump, hate_crimes$avg_hatecrimes_per_100k_fbi)
```

```
## [1] -0.5580764
```

```
# controlling for Trump votes
voted_trump <- hate_crimes %>%
  group_by(state) %>%
  select(share_voters_voted_trump, share_white_poverty, avg_hatecrimes_per_100k_fbi) %>%
              arrange(desc(share_voters_voted_trump))
```

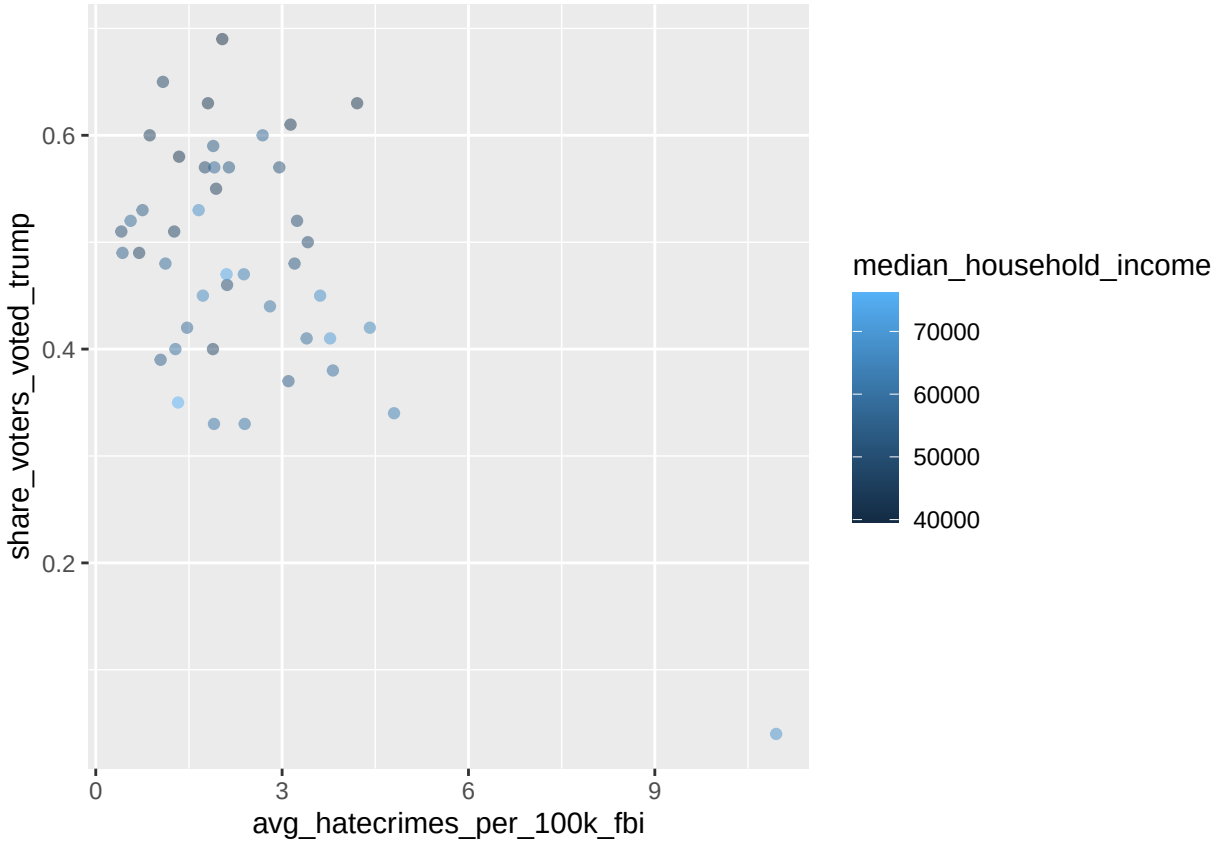
```
## Adding missing grouping variables: `state`
```

```
 voted_trump
```

```
## # A tibble: 45 x 4
## # Groups:   state [45]
##    state       share_voters_voted_t~1 share_white_poverty avg_hatecrimes_per_1~2
##    <chr>                        <dbl>               <dbl>                  <dbl>
##  1 West Virgi~                   0.69                0.14                   2.04
##  2 Oklahoma                      0.65                0.1                    1.08
##  3 Alabama                       0.63                0.12                   1.81
##  4 Kentucky                      0.63                0.17                   4.21
##  5 Tennessee                     0.61                0.13                   3.14
##  6 Arkansas                      0.6                 0.12                   0.869
##  7 Nebraska                      0.6                 0.07                   2.69
##  8 Idaho                         0.59                0.11                   1.89
##  9 Louisiana                     0.58                0.12                   1.34
```

```
## 10 Indiana                        0.57                0.12                  1.76
## # i 35 more rows
## # i abbreviated names: 1: share_voters_voted_trump,
## #   2: avg_hatecrimes_per_100k_fbi
```

```
ggplot(hate_crimes, aes(x =avg_hatecrimes_per_100k_fbi, y = share_voters_voted_trump, color = median_hou
  geom_point(alpha = 0.5)
```




## Findings and Conclusion

None of the variables taken alone fully explains the average hate crime rates noticed in the data frame. This could point to the explanation that these crimes are the result of a combination of factors. Geographically, the states with the lowest average hate crime rates are typically not "border" states. Many of them of these states are located in the U.S. hinterland and don't have/share an international border. These states are:
Illinois
Arkansas
Texas
Florida Iowa
Pennsylvania
Georgia

Likewise, states that experience the highest hate crime are typically not the ones with lowest household income range, nor are they "border" states. Besides New Jersey, many of these state are also located in the U.S hinterland. And, except for Kentucky, many of these states are affluent states, with substantial size of the population holding high school degrees. These states are in descending order: District of Columbia
Massachusetts
New Jersey
Kentucky

Washington
Connecticut Minnesota Finally, voting for Trump was not found to increase or decrease hate crimes across states.