DATA2020FinalProject_plan

Hanjun Wei, Keying Gong, Yurui Zhang

4/26/2022

Data

The data our group uses is the Police Shootings in the US: https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

Question

Our target question: What will the total number of fatal shootings by on duty police officers all over the united states next week?

Our brief plan

The objective we are interested in: 1. Analyze historical fatal shootings based on demographic groups 2. predict 2022 fatal shootings using time-series model for the whole population and for different racial groups

Step 1: data cleaning and conduct feature engineering

Step 2: Exploratory Data Analysis on the police shooting dataset. Discover interesting patterns

Step 3: Compare race composition of each state VS. fatal shootings race composition by state

Step 4: (Time Series Modeling) build ARIMA model to predict 2022 fatal shootings using time-series model for the whole population and for different racial groups

(Racial Mixed Effect Modeling) Reconstruct the dataset in terms of Race.

Feature #1 is Racial Category, which includes 7 types of race.

Feature #2 is Number of Fatal Shooting at each time stamp, count the total number of fatal shootings in a given time interval.

Feature #3 is Time Stamp, month and year. Construct a mixed effect modeling in terms of the number of fatal shootings.

(State Mixed Effect Modeling) Reconstruct the dataset in terms of State.

Feature #1 is State Category, which includes 50 types of state.

Feature #2 is Number of Fatal Shooting at each time stamp, count the total number of fatal shootings in a given time interval.

Feature #3 is Time Stamp, month and year. Construct a mixed effect modeling in terms of the number of fatal shootings.

Data Cleaning first

First thing first, read the data into the file.

Print out the first 10 rows to better understand the data.

```
## # A tibble: 10 x 17
##       id name    date       manner_of_death armed   age gender race  city  state
##    <dbl> <chr>   <date>     <chr>           <chr> <dbl> <chr>  <chr> <chr> <chr>
##  1     3 Tim El~ 2015-01-02 shot            gun      53 M      A     Shel~ WA
##  2     4 Lewis ~ 2015-01-02 shot            gun      47 M      W     Aloha OR
##  3     5 John P~ 2015-01-03 shot and Taser~ unar~    23 M      H     Wich~ KS
##  4     8 Matthe~ 2015-01-04 shot            toy ~    32 M      W     San ~ CA
##  5     9 Michae~ 2015-01-04 shot            nail~    39 M      H     Evans CO
##  6    11 Kennet~ 2015-01-04 shot            gun      18 M      W     Guth~ OK
##  7    13 Kennet~ 2015-01-05 shot            gun      22 M      H     Chan~ AZ
##  8    15 Brock ~ 2015-01-06 shot            gun      35 M      W     Assa~ KS
##  9    16 Autumn~ 2015-01-06 shot            unar~    34 F      W     Burl~ IA
## 10    17 Leslie~ 2015-01-06 shot            toy ~    47 M      B     Knox~ PA
## # ... with 7 more variables: signs_of_mental_illness <lgl>, threat_level <chr>,
## #   flee <chr>, body_camera <lgl>, longitude <dbl>, latitude <dbl>,
## #   is_geocoding_exact <lgl>
```

Looks good to me.

let us print the names of all of the columns to see if it matches our key table.

```
##  [1] "id"                     "name"
##  [3] "date"                   "manner_of_death"
##  [5] "armed"                  "age"
##  [7] "gender"                 "race"
##  [9] "city"                   "state"
## [11] "signs_of_mental_illness" "threat_level"
## [13] "flee"                   "body_camera"
## [15] "longitude"              "latitude"
## [17] "is_geocoding_exact"
```

Then after we read in the data, we can take a look at some characteristic of the data, First of all, we print out the shape of the data.

```
## [1] 7291   17
```

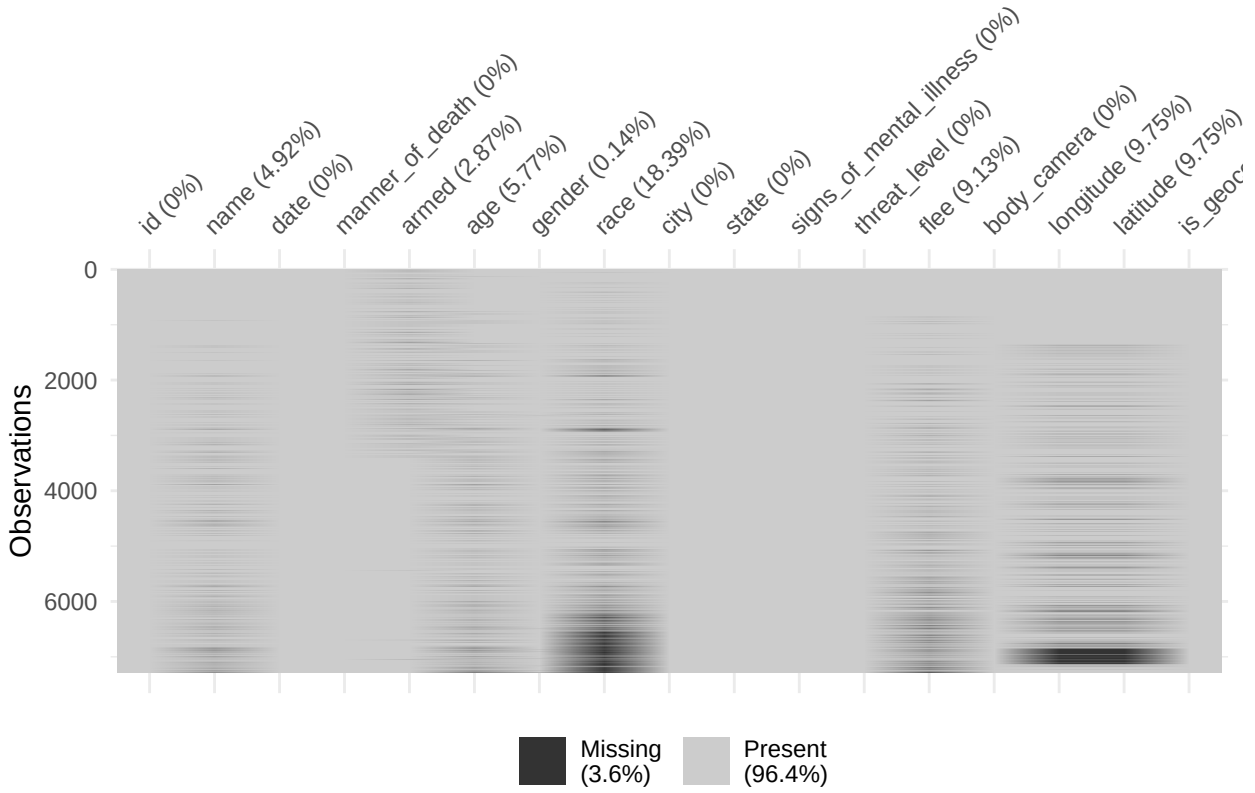
We have 7291 rows and 17 columns.

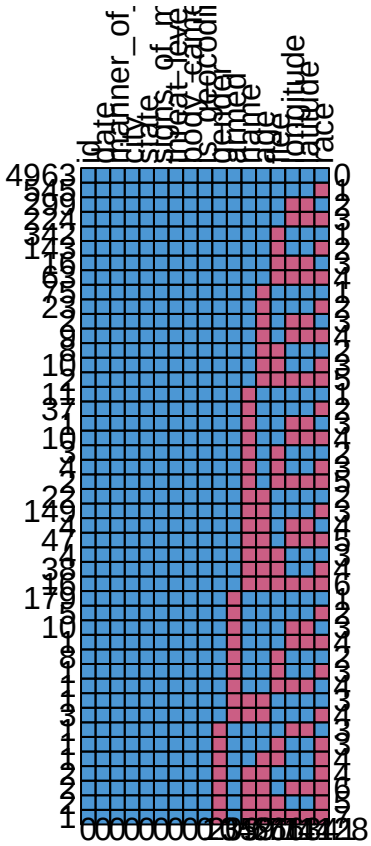
we want to know the exact percentage of missing value in each columns, so we print out the percentage of not missing values in our datasets.

```
##                      id                    name                    date
##               1.0000000               0.9507612               1.0000000
##         manner_of_death                   armed                     age
##               1.0000000               0.9713345               0.9422576
##                  gender                    race                    city
##               0.9986284               0.8160746               1.0000000
##                   state signs_of_mental_illness            threat_level
##               1.0000000               1.0000000               1.0000000
##                    flee             body_camera               longitude
##               0.9086545               1.0000000               0.9024825
##                latitude      is_geocoding_exact
##               0.9024825               1.0000000
```

From the plot, we can see that there are missing values in name, armed, age, gender, race, flee, longitude, and latitude.

Then we will run the missing pattern plots to take a better look at the missineness.





```
##      id date manner_of_death city state signs_of_mental_illness threat_level
## 4963  1    1               1    1     1                       1            1
## 545   1    1               1    1     1                       1            1
## 299   1    1               1    1     1                       1            1
## 224   1    1               1    1     1                       1            1
## 342   1    1               1    1     1                       1            1
## 143   1    1               1    1     1                       1            1
## 16    1    1               1    1     1                       1            1
## 63    1    1               1    1     1                       1            1
## 75    1    1               1    1     1                       1            1
## 23    1    1               1    1     1                       1            1
## 2     1    1               1    1     1                       1            1
## 9     1    1               1    1     1                       1            1
## 8     1    1               1    1     1                       1            1
## 10    1    1               1    1     1                       1            1
## 2     1    1               1    1     1                       1            1
## 11    1    1               1    1     1                       1            1
## 37    1    1               1    1     1                       1            1
## 1     1    1               1    1     1                       1            1
## 10    1    1               1    1     1                       1            1
## 3     1    1               1    1     1                       1            1
## 4     1    1               1    1     1                       1            1
## 2     1    1               1    1     1                       1            1
## 22    1    1               1    1     1                       1            1
## 149   1    1               1    1     1                       1            1
## 4     1    1               1    1     1                       1            1
## 47    1    1               1    1     1                       1            1
```

```
## 4     1    1                1    1     1                       1             1
## 38    1    1                1    1     1                       1             1
## 16    1    1                1    1     1                       1             1
## 179   1    1                1    1     1                       1             1
## 5     1    1                1    1     1                       1             1
## 10    1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 8     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 3     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
## 2     1    1                1    1     1                       1             1
## 2     1    1                1    1     1                       1             1
## 2     1    1                1    1     1                       1             1
## 1     1    1                1    1     1                       1             1
##       0    0                0    0     0                       0             0
##      body_camera is_geocoding_exact gender armed name age flee longitude
## 4963           1                  1      1     1    1   1    1         1
## 545            1                  1      1     1    1   1    1         1
## 299            1                  1      1     1    1   1    1         0
## 224            1                  1      1     1    1   1    1         0
## 342            1                  1      1     1    1   1    0         1
## 143            1                  1      1     1    1   1    0         1
## 16             1                  1      1     1    1   1    0         0
## 63             1                  1      1     1    1   1    0         0
## 75             1                  1      1     1    1   0    1         1
## 23             1                  1      1     1    1   0    1         1
## 2              1                  1      1     1    1   0    1         0
## 9              1                  1      1     1    1   0    1         0
## 8              1                  1      1     1    1   0    0         1
## 10             1                  1      1     1    1   0    0         1
## 2              1                  1      1     1    1   0    0         0
## 11             1                  1      1     1    0   1    1         1
## 37             1                  1      1     1    0   1    1         1
## 1              1                  1      1     1    0   1    1         0
## 10             1                  1      1     1    0   1    1         0
## 3              1                  1      1     1    0   1    0         1
## 4              1                  1      1     1    0   1    0         1
## 2              1                  1      1     1    0   1    0         0
## 22             1                  1      1     1    0   0    1         1
## 149            1                  1      1     1    0   0    1         1
## 4              1                  1      1     1    0   0    1         0
## 47             1                  1      1     1    0   0    1         0
## 4              1                  1      1     1    0   0    0         1
## 38             1                  1      1     1    0   0    0         1
## 16             1                  1      1     1    0   0    0         0
## 179            1                  1      1     0    1   1    1         1
## 5              1                  1      1     0    1   1    1         1
## 10             1                  1      1     0    1   1    1         0
## 1              1                  1      1     0    1   1    1         0
```

```
## 8              1                  1      1     0    1   1    0          1
## 1              1                  1      1     0    1   1    0          1
## 1              1                  1      1     0    1   1    0          0
## 1              1                  1      1     0    0   0    1          1
## 3              1                  1      1     0    0   0    1          1
## 1              1                  1      0     1    1   1    1          0
## 1              1                  1      0     1    1   1    0          1
## 1              1                  1      0     1    1   0    0          1
## 2              1                  1      0     1    0   0    1          1
## 2              1                  1      0     1    0   0    1          0
## 2              1                  1      0     1    0   0    0          1
## 1              1                  1      0     1    0   0    0          0
##                0                  0     10   209  359 421  666        711
##      latitude race
## 4963        1    1    0
## 545         1    0    1
## 299         0    1    2
## 224         0    0    3
## 342         1    1    1
## 143         1    0    2
## 16          0    1    3
## 63          0    0    4
## 75          1    1    1
## 23          1    0    2
## 2           0    1    3
## 9           0    0    4
## 8           1    1    2
## 10          1    0    3
## 2           0    0    5
## 11          1    1    1
## 37          1    0    2
## 1           0    1    3
## 10          0    0    4
## 3           1    1    2
## 4           1    0    3
## 2           0    0    5
## 22          1    1    2
## 149         1    0    3
## 4           0    1    4
## 47          0    0    5
## 4           1    1    3
## 38          1    0    4
## 16          0    0    6
## 179         1    1    1
## 5           1    0    2
## 10          0    1    3
## 1           0    0    4
## 8           1    1    2
## 1           1    0    3
## 1           0    1    4
## 1           1    1    3
## 3           1    0    4
## 1           0    1    3
## 1           1    0    3
```

```
## 1           1    0    4
## 2           1    0    4
## 2           0    0    6
## 2           1    0    5
## 1           0    0    7
##           711 1341 4428
```

For our variable “id”, from the code book, this should be the continuous variable that contains the id information.

```
## [1] 7291
```

For our variable “name”, from the code book, this should be the categorical variable that contains the id information.

```
## [1] 6902
```

For our variable “date”, from the code book, this should be the date variable. So we add the Year, Month and WeekDay new variables.

```
## [1] 2485
```

For our variable “manner_of_death”, from the code book, this should be the categorical variable.

```
## [1] "shot"             "shot and Tasered"
```

For our variable “armed”, from the code book, this should be the categorical variable.

For our variable “age”, from the code book, this should be the categorical variable.

```
##  [1] 53 47 23 32 39 18 22 35 34 25 31 41 30 37 28 42 36 49 71 33 29 43 24 75 68
## [26] 27 48 21 67 19 54 17 56 61 45 26 40 59 38 51 74 57 46 16 50 20 77 NA 58 64
## [51] 52 63 44 60 66 83 72 76 62 55 69 86 15 65  6 12 70 80 14 82 13 73 91 79 78
## [76] 84 81 89 88  8 92
```

```
## [1] "(35-55) middle age"      "(18-35) young adulthood"
## [3] "(0-18) pre-young"        "(>55) older adulthood"
## [5] NA
```

For our variable “gender”, from the code book, this should be the categorical variable.

```
## [1] "M" "F" NA
```

For our variable “race”, from the code book, this should be the categorical variable.

```
## [1] "A" "W" "H" "B" "O" NA  "N"
```

For our variable “city”, from the code book, this should be the categorical variable.

```
## [1] 3032
```

For our variable “state”, from the code book, this should be the categorical variable.

```
##  [1] "WA" "OR" "KS" "CA" "CO" "OK" "AZ" "IA" "PA" "TX" "OH" "LA" "MT" "UT" "AR"
## [16] "IL" "NV" "NM" "MN" "MO" "VA" "NJ" "IN" "KY" "MA" "NH" "FL" "ID" "MD" "NE"
## [31] "MI" "GA" "TN" "NC" "AK" "NY" "ME" "AL" "MS" "WI" "SC" "DE" "DC" "WV" "HI"
## [46] "WY" "ND" "CT" "SD" "VT" "RI"
```

```
## [1] "the_west"   "mid_west"   "south_west" "north_east" "south_east"
```

For our variable “signs_of_mental_illness”, from the code book, this should be the categorical variable.

```
## [1]  TRUE FALSE
```

For our variable “threat_level”, from the code book, this should be the categorical variable.

```
## [1] "attack"       "other"        "undetermined"
```

For our variable "flee", from the code book, this should be the categorical variable.

```
## [1] "Not fleeing" "Car"         "Foot"        "Other"       NA
```

For our variable "body_camera", from the code book, this should be the categorical variable.

```
## [1] FALSE  TRUE
```

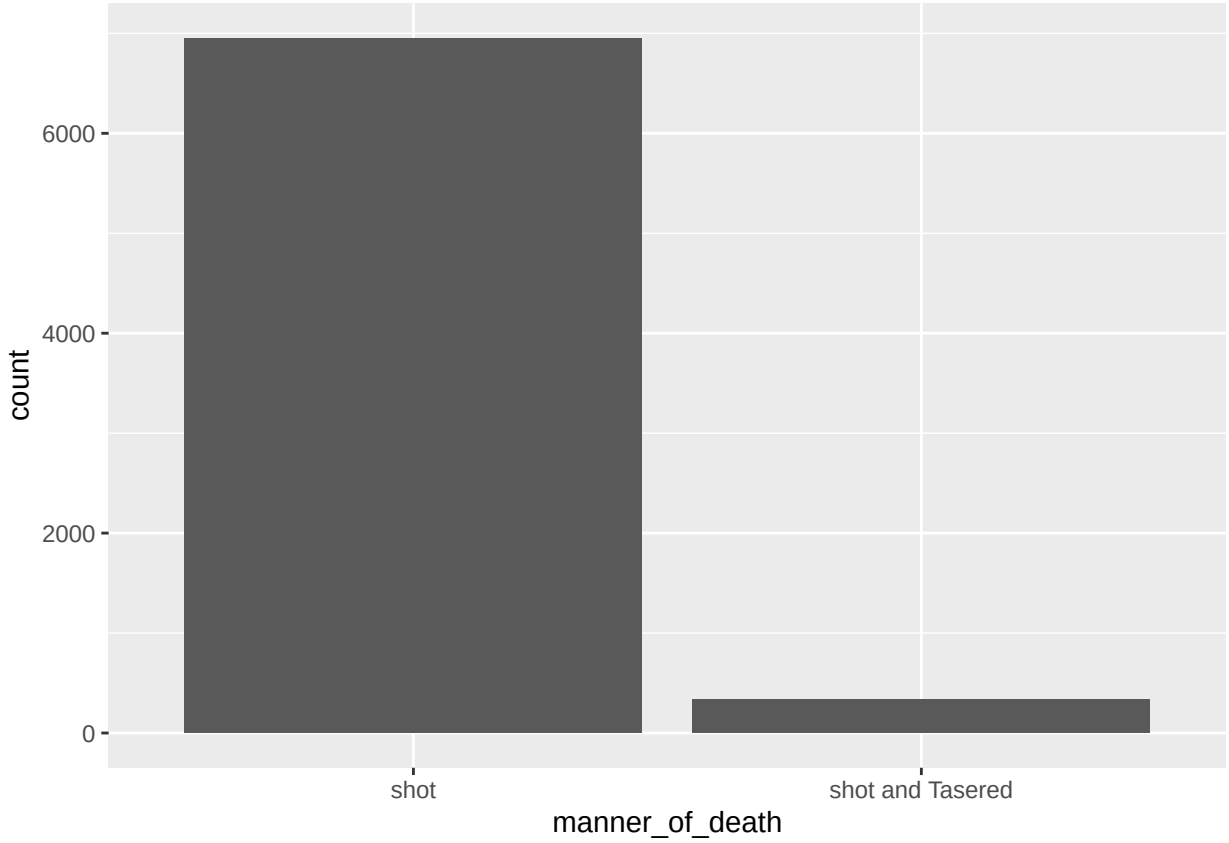
Now we have to create two new datasets, one for eda and analysis, one for creating model.

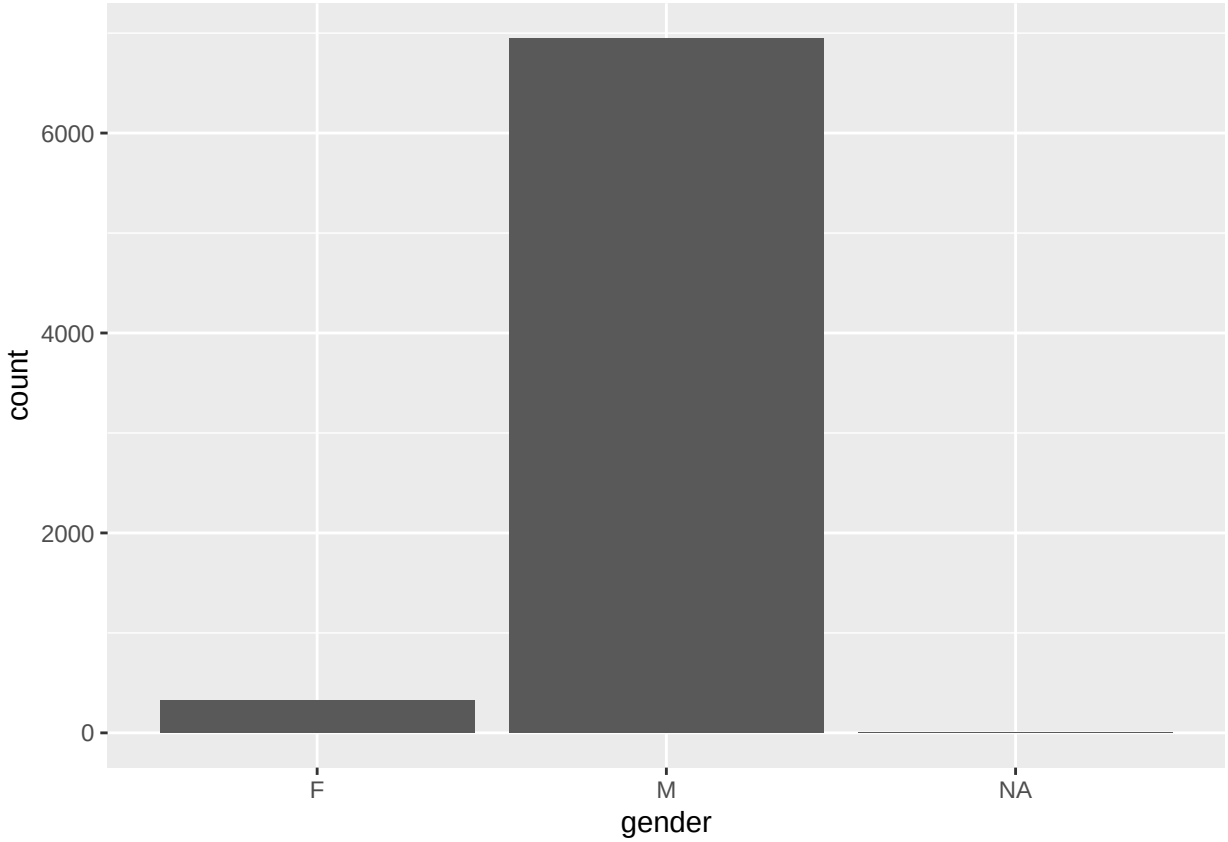
```
##  [1] "id"                     "name"
##  [3] "date"                   "manner_of_death"
##  [5] "armed"                  "age"
##  [7] "gender"                 "race"
##  [9] "city"                   "state"
## [11] "signs_of_mental_illness" "threat_level"
## [13] "flee"                   "body_camera"
## [15] "longitude"              "latitude"
## [17] "is_geocoding_exact"     "Year"
## [19] "Month"                  "WeekDay"
## [21] "armed_level"            "age_group"
## [23] "state_loc"
```

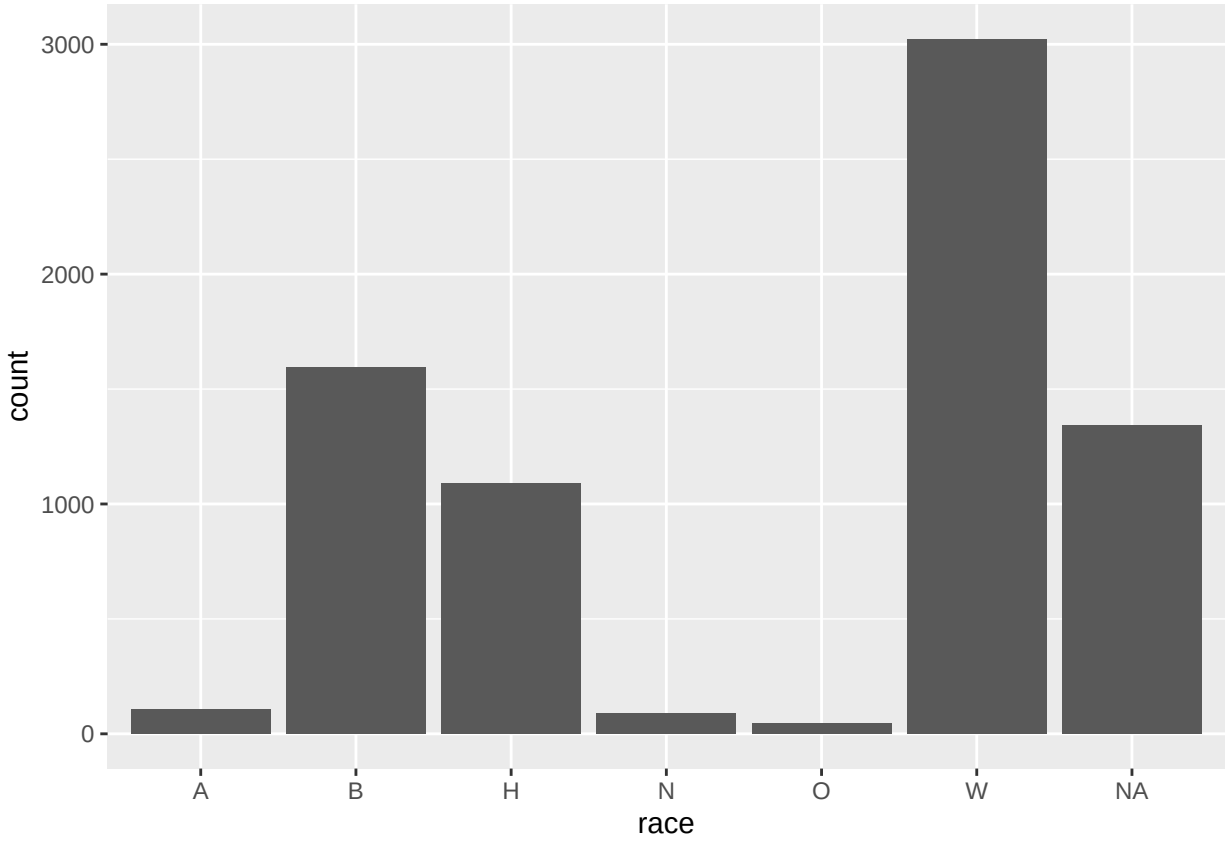
EDA

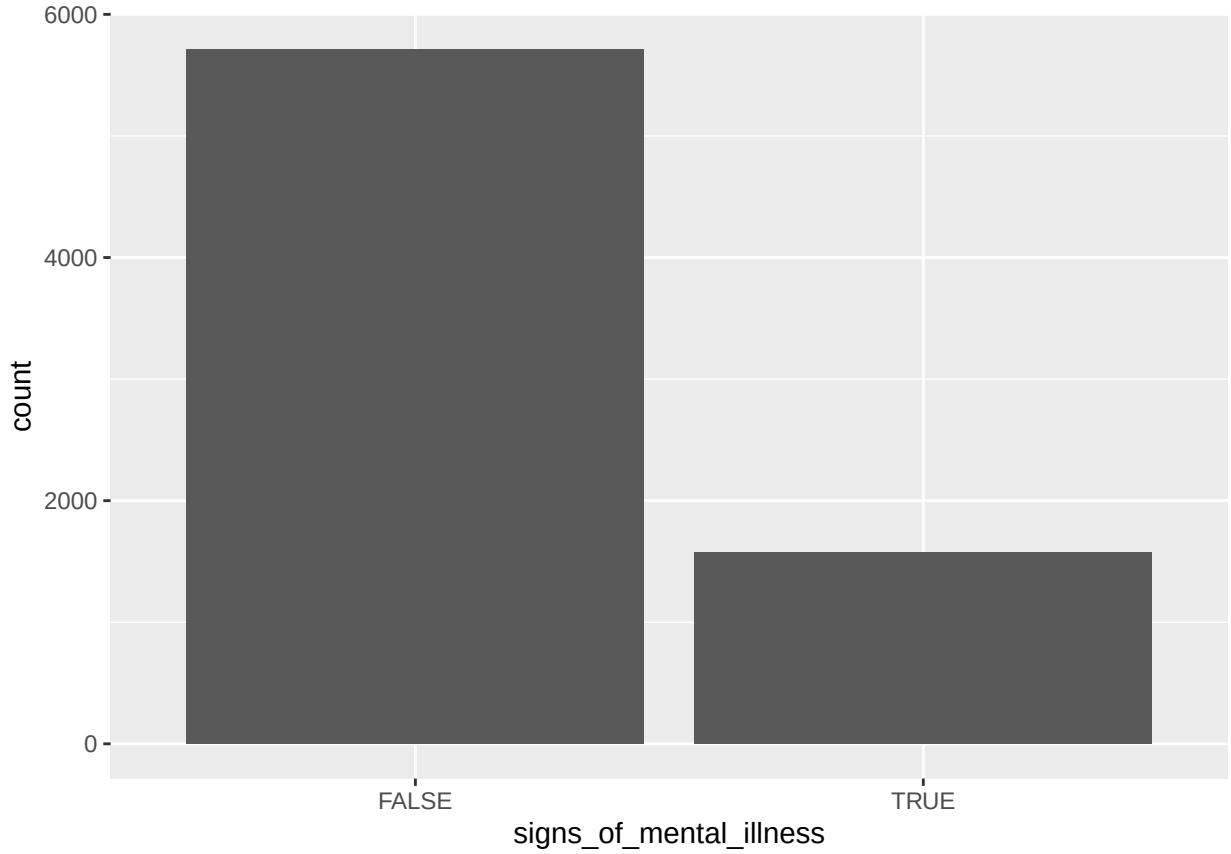
```
##        id           name               date                     manner_of_death
##  Min.   :   3   Length:7291        Min.   :2015-01-02   shot            :6952
##  1st Qu.:2046   Class :character   1st Qu.:2016-11-08   shot and Tasered: 339
##  Median :4051   Mode  :character   Median :2018-09-04
##  Mean   :4032                      Mean   :2018-09-08
##  3rd Qu.:6010                      3rd Qu.:2020-07-12
##  Max.   :7951                      Max.   :2022-04-20
##
##     armed                age          gender       race          city
##  Length:7291        Min.   : 6.00   F   : 330   A   : 105   Length:7291
##  Class :character   1st Qu.:27.00   M   :6951   B   :1595   Class :character
##  Mode  :character   Median :35.00   NA's:  10   H   :1089   Mode  :character
##                     Mean   :37.15               N   :  91
##                     3rd Qu.:45.00               O   :  47
##                     Max.   :92.00               W   :3023
##                     NA's   :421                 NA's:1341
##      state      signs_of_mental_illness       threat_level          flee
##  CA     :1063   FALSE:5715              attack      :4667   Car        :1183
##  TX     : 642   TRUE :1576              other       :2364   Foot       : 943
##  FL     : 461                           undetermined: 260   Not fleeing:4232
##  AZ     : 334                                               Other      : 267
##  GA     : 272                                               NA's       : 666
##  CO     : 262
##  (Other):4257
##  body_camera    longitude          latitude     is_geocoding_exact
##  FALSE:6272   Min.   :-160.01   Min.   :19.50   Mode :logical
##  TRUE :1019   1st Qu.:-112.07   1st Qu.:33.48   FALSE:18
##               Median : -94.26   Median :36.08   TRUE :7273
##               Mean   : -97.10   Mean   :36.66
##               3rd Qu.: -83.16   3rd Qu.:40.00
```

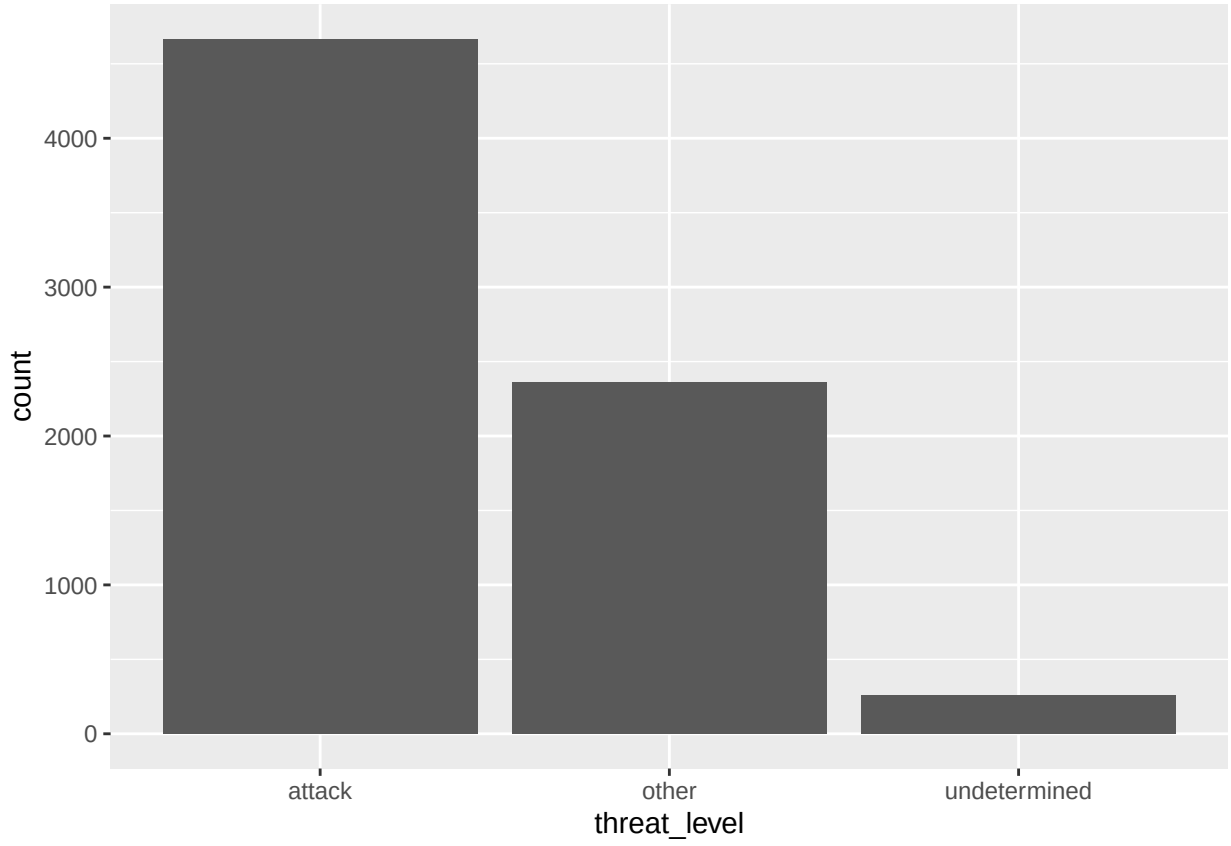
```
##               Max.   : -67.87   Max.   :71.30
##               NA's   :711       NA's   :711
##       Year          Month           WeekDay                 armed_level
##  2021   :1054   3      : 728   Friday   :1041   long_dis_death    :4258
##  2020   :1020   1      : 675   Monday   : 993   short_dis_death   :1241
##  2019   : 999   2      : 672   Saturday : 962   unarmed           : 443
##  2015   : 994   4      : 615   Sunday   :1000   undetermined      : 365
##  2018   : 985   10     : 604   Thursday :1087   long_dis_not_death: 304
##  2017   : 981   8      : 602   Tuesday  :1108   (Other)           : 471
##  (Other):1258   (Other):3395   Wednesday:1100   NA's              : 209
##                    age_group        state_loc
##  (>55) older adulthood  : 701   mid_west  :1178
##  (0-18) pre-young       : 241   north_east: 509
##  (18-35) young adulthood:3333   south_east:2207
##  (35-55) middle age     :2595   south_west:1337
##  NA's                   : 421   the_west  :2060
##
##
```

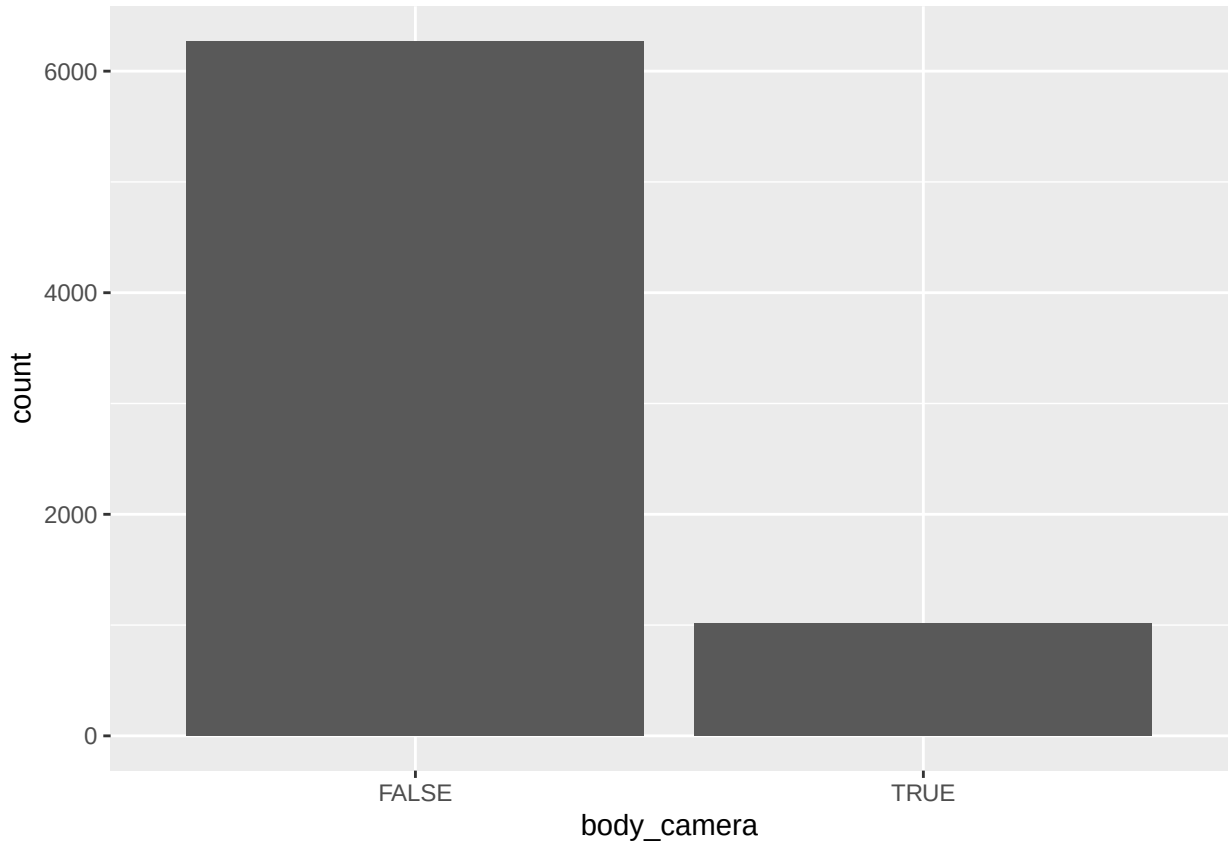


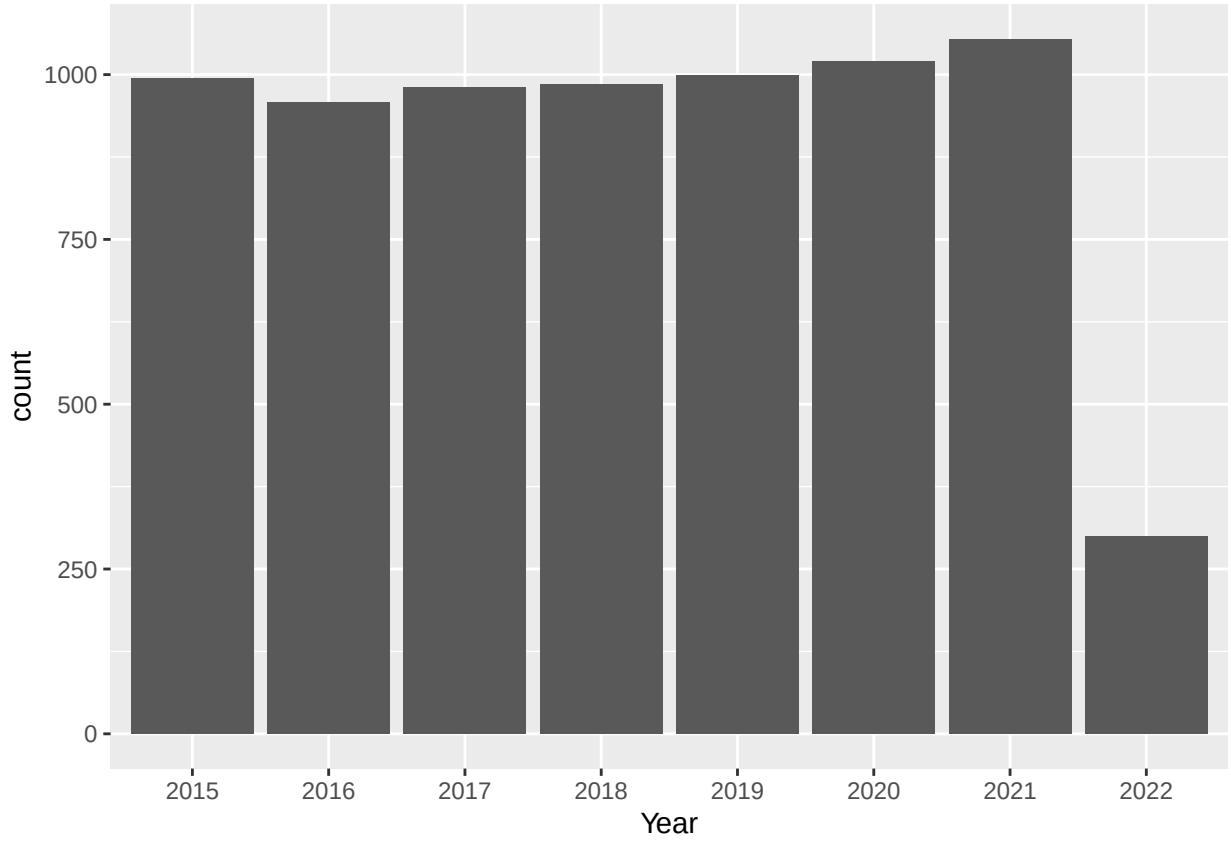

6000
4000
2000
0
count
F
M
NA
gender

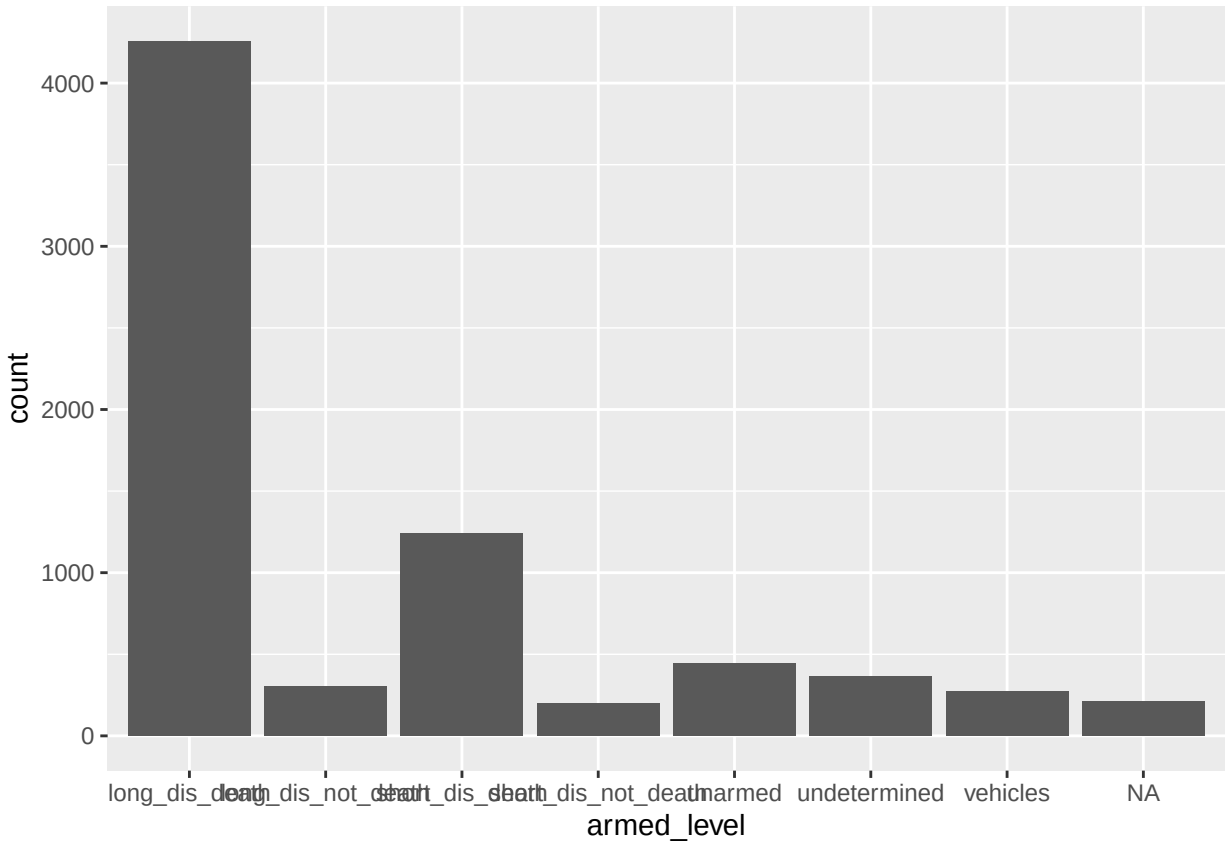

3000
2000
1000
0
count
A
B
H
N
O
W
NA
race

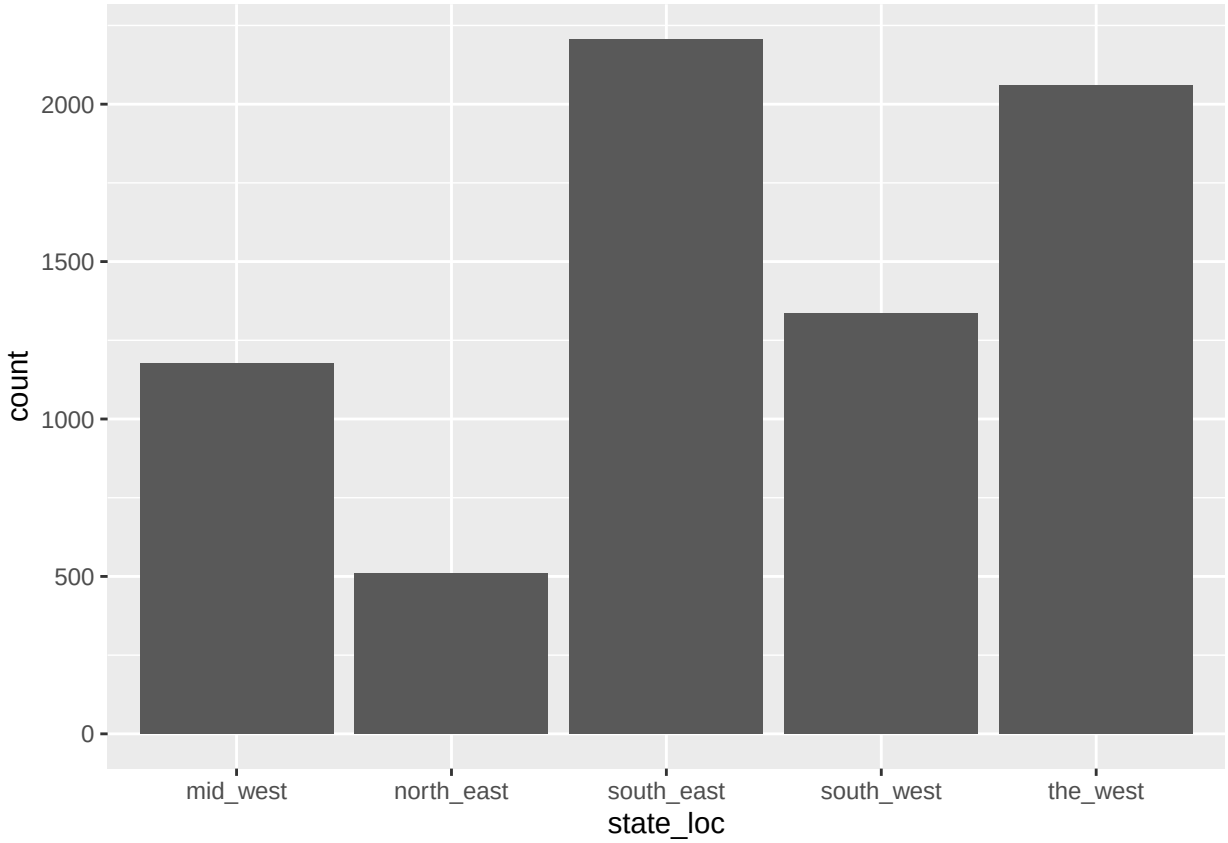

6000
4000
2000
0
count
FALSE
TRUE
signs_of_mental_illness

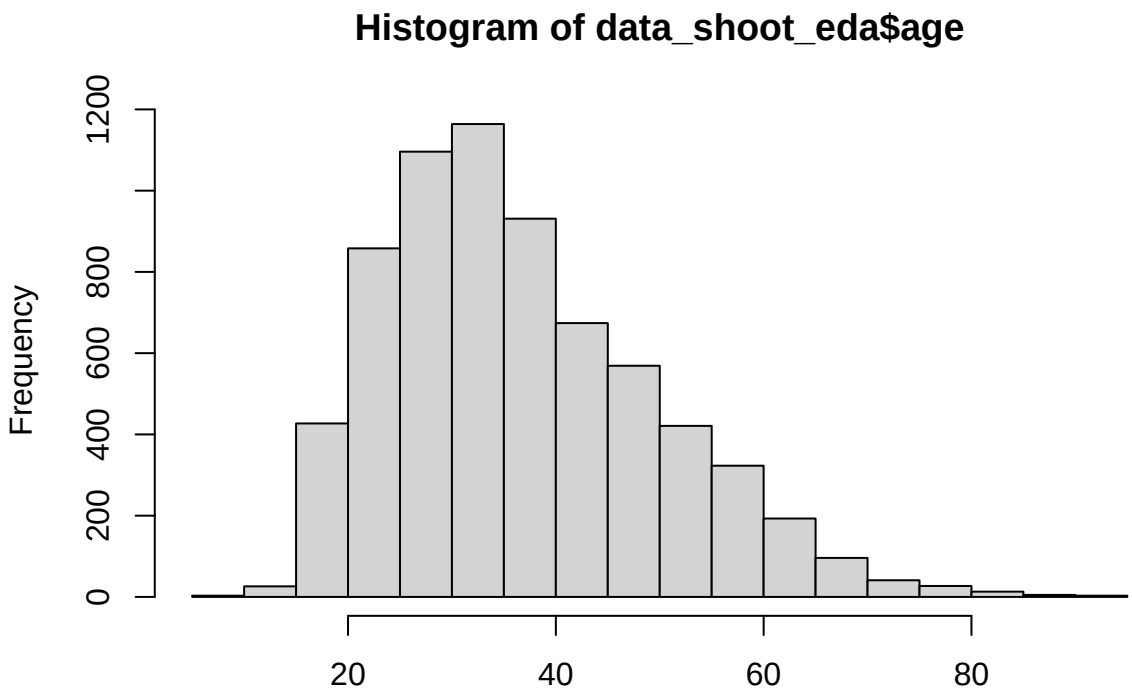

4000
3000
2000
1000
0
count
attack
other
undetermined
threat_level

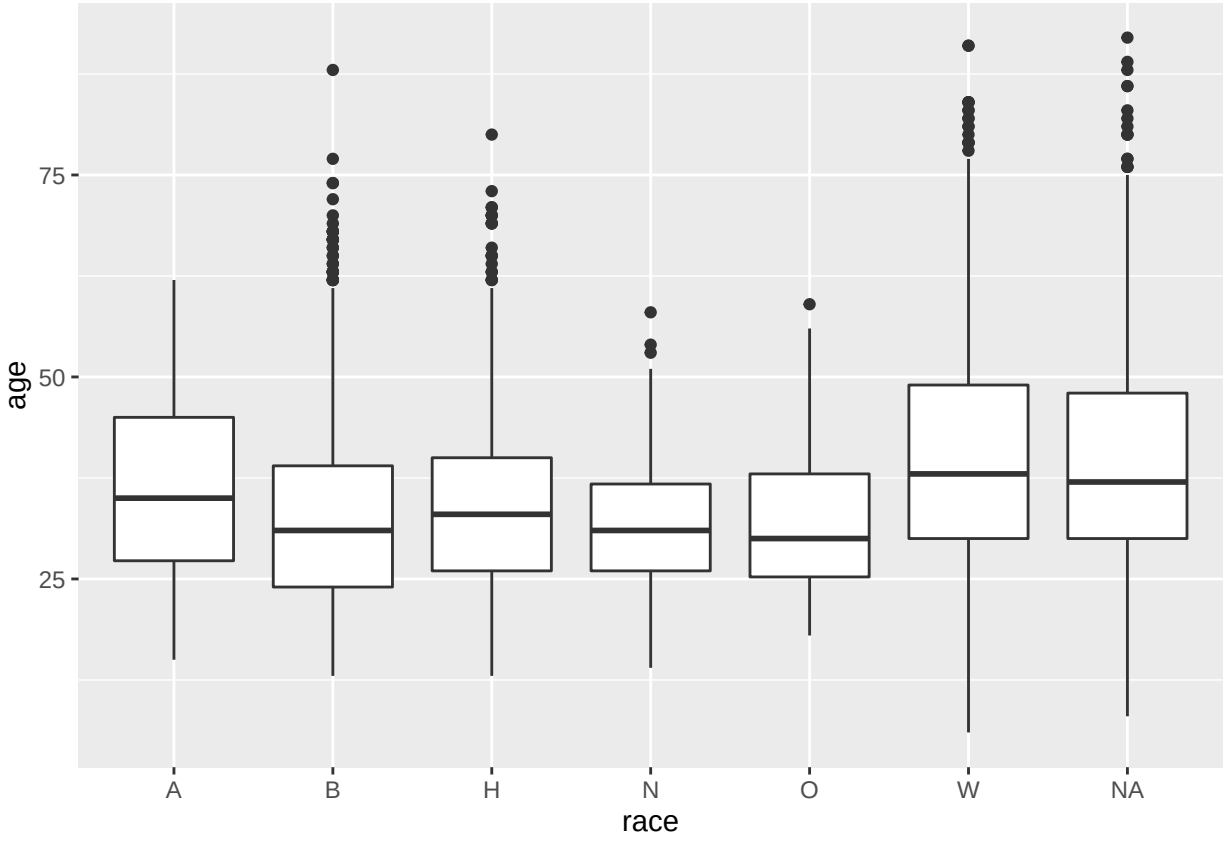

6000
4000
2000
0
count
FALSE
TRUE
body_camera


1000
750
500
250
0
count
2015
2016
2017
2018
2019
2020
2021
2022
Year


4000
3000
2000
1000
0
count
long_dis_death
long_dis_not_death
short_dis_death
short_dis_not_death
unarmed
undetermined
vehicles
NA
armed_level


2000
1500
1000
500
0
count
mid_west
north_east
south_east
south_west
the_west
state_loc






```
geo(address = c("Tokyo", "Lima", "Nairobi"),
 method = 'osm')
```

```
## Passing 3 addresses to the Nominatim single address geocoder
## Query completed in: 3 seconds
## # A tibble: 3 x 3
##   address    lat  long
##   <chr>    <dbl> <dbl>
## 1 Tokyo     35.7  140.
## 2 Lima     -12.1  -77.0
## 3 Nairobi  -1.28  36.8
```