Carreguem les llibreries que utilitzarem

```r
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
```

```r
library(ggplot2)
library(magrittr)
```

```r
df_householdIncome <- read.csv("data/MedianHouseholdIncome2015.csv", sep=",")
df_poverty <- read.csv("data/PercentagePeopleBelowPovertyLevel.csv", sep=",")
df_highSchool <- read.csv("data/PercentOver25CompletedHighSchool.csv", sep=",")
df_policeKilling <- read.csv("data/PoliceKillingsUS.csv", sep=",")
df_shareRace <- read.csv("data/ShareRaceByCity.csv", sep=",")

head(df_householdIncome)
```

```
##   Geographic.Area            City Median.Income
## 1              AL      Abanda CDP         11207
## 2              AL  Abbeville city         25615
## 3              AL Adamsville city         42575
## 4              AL    Addison town         37083
## 5              AL      Akron town         21667
## 6              AL  Alabaster city         71816
```

```r
head(df_poverty)
```

```
##   Geographic.Area            City poverty_rate
## 1              AL      Abanda CDP         78.8
## 2              AL  Abbeville city         29.1
## 3              AL Adamsville city         25.5
## 4              AL    Addison town         30.7
## 5              AL      Akron town           42
## 6              AL  Alabaster city         11.2
```

```r
head(df_highSchool)
```

```
##   Geographic.Area            City percent_completed_hs
## 1              AL      Abanda CDP                 21.2
## 2              AL  Abbeville city                 69.1
## 3              AL Adamsville city                 78.9
## 4              AL    Addison town                 81.4
## 5              AL      Akron town                 68.6
## 6              AL  Alabaster city                 89.3
```

```r
head(df_policeKilling)
```

```
##   id               name     date  manner_of_death      armed age gender race
## 1  3         Tim Elliot 02/01/15             shot        gun  53      M    A
## 2  4   Lewis Lee Lembke 02/01/15             shot        gun  47      M    W
## 3  5 John Paul Quintero 03/01/15 shot and Tasered    unarmed  23      M    H
## 4  8    Matthew Hoffman 04/01/15             shot toy weapon  32      M    W
## 5  9  Michael Rodriguez 04/01/15             shot   nail gun  39      M    H
## 6 11  Kenneth Joe Brown 04/01/15             shot        gun  18      M    W
##            city state signs_of_mental_illness threat_level        flee
## 1       Shelton    WA                    TRUE       attack Not fleeing
## 2         Aloha    OR                   FALSE       attack Not fleeing
## 3       Wichita    KS                   FALSE        other Not fleeing
## 4 San Francisco    CA                    TRUE       attack Not fleeing
## 5         Evans    CO                   FALSE       attack Not fleeing
## 6       Guthrie    OK                   FALSE       attack Not fleeing
##   body_camera
## 1       FALSE
## 2       FALSE
## 3       FALSE
## 4       FALSE
## 5       FALSE
## 6       FALSE
```

```r
head(df_shareRace)
```

```
##   Geographic.area            City share_white share_black share_native_american
## 1              AL      Abanda CDP        67.2        30.2                     0
## 2              AL  Abbeville city        54.4        41.4                   0.1
## 3              AL Adamsville city        52.3        44.9                   0.5
## 4              AL    Addison town        99.1         0.1                     0
## 5              AL      Akron town        13.2        86.5                     0
## 6              AL  Alabaster city        79.4        13.5                   0.4
##   share_asian share_hispanic
## 1           0            1.6
## 2           1            3.1
## 3         0.3            2.3
## 4         0.1            0.4
## 5           0            0.3
## 6         0.9              9
```

Canviem el nom de les columnes

```r
colnames(df_householdIncome)[1] <- "area_geografica"
colnames(df_poverty)[1] <- "area_geografica"
colnames(df_highSchool)[1] <- "area_geografica"
colnames(df_shareRace)[1] <- "area_geografica"
```

Merge els distins df:

```
USAv1 <- merge(df_highSchool, df_poverty, by.x=c("area_geografica", "City"), by.y=c("area_geografica",
USAv2 <- merge(USAv1, df_householdIncome, by.x=c("area_geografica", "City"), by.y=c("area_geografica",
USA <- merge(USAv2, df_shareRace, by.x=c("area_geografica", "City"), by.y=c("area_geografica", "City"))
```

Normalitzem els noms de les ciutats:

```
USA$City <- gsub(" CDP| city| town|\\.| ","", USA$City)
df_policeKilling$city <- gsub(" County| Parish|[^[:alnum:]]","",df_policeKilling$city)
```

Merge del dataframe ambtingut amb df_policeKilling i neteja i preparació de les dades:

```
df_clean <- merge(df_policeKilling, USA, by.x=c("state", "city"), by.y=c("area_geografica", "City"))

#df_clean$id <- NULL
#df_clean$city <- NULL
#df_clean$state <- NULL

# Convertim el camp date de tipus character a tipus date
df_clean %<>% mutate(date=as.Date(date, format = "%d/%m/%y"))

#rownames(df_clean) <- 1:nrow(df_clean)
```

Tractar camp Median.Income:

```
table(df_clean$Median.Income)[1:5]
```

```
##
##      -    (X) 100469 100849 101689
##      1      6      1      1      1
```

```
# Hem vist que la variable Median.Income te el valor "-" i "(X)", els subtituim per 0
df_clean[df_clean$Median.Income == "-",]$Median.Income <- "0"
df_clean[df_clean$Median.Income == "(X)",]$Median.Income <- "0"

# Convertim la variable a tipus numeric
df_clean$Median.Income <- as.numeric(df_clean$Median.Income)

# Calculem la mitjana i la asignem als valors que haviem subtituit abans
mean_income <- mean(df_clean[df_clean$Median.Income > 0,]$Median.Income)
df_clean$Median.Income[df_clean$Median.Income == 0] <- mean_income
```

Continuem amb el tractament de les dades:

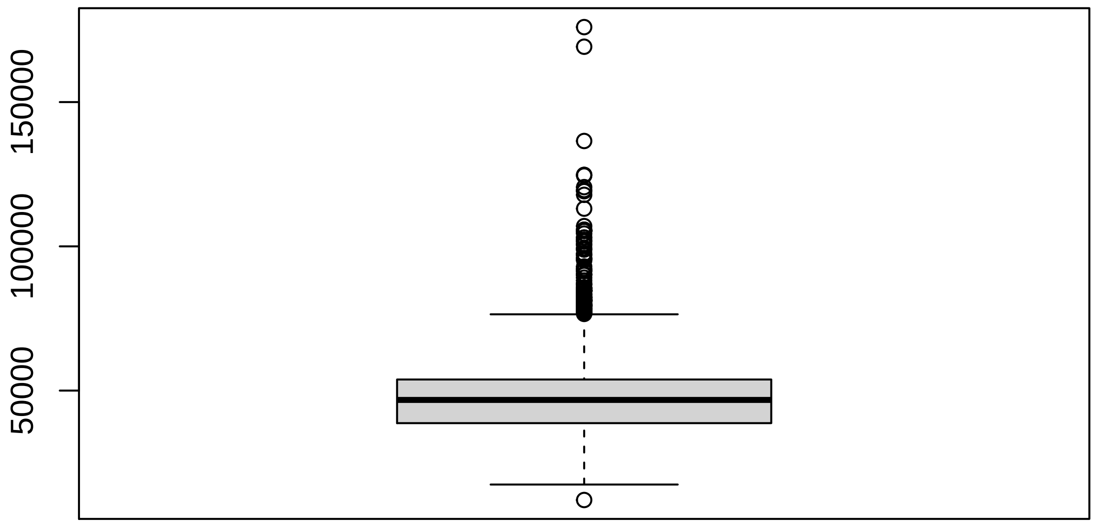
- Pasarem les variables: manner_of_death, armed, gender, race, threat_level i flee a tipus factor.
- I les variables: percent_completed_hs, poverty_rate, share_white, share_asian, share_black, share_native_american i share_hispanic a tipus numeric.

```
df_clean$manner_of_death <- as.factor(df_clean$manner_of_death)
df_clean$armed <- as.factor(df_clean$armed)
df_clean$gender <- as.factor(df_clean$gender)
```

```
df_clean$race <- as.factor(df_clean$race)
df_clean$threat_level <- as.factor(df_clean$threat_level)
df_clean$flee <- as.factor(df_clean$flee)
df_clean$percent_completed_hs <- as.numeric(df_clean$percent_completed_hs)
df_clean$poverty_rate <- as.numeric(df_clean$poverty_rate)
df_clean$share_white <- as.numeric(df_clean$share_white)
df_clean$share_asian <- as.numeric(df_clean$share_asian)
df_clean$share_black <- as.numeric(df_clean$share_black)
df_clean$share_native_american <- as.numeric(df_clean$share_native_american)
df_clean$share_hispanic <- as.numeric(df_clean$share_hispanic)
```

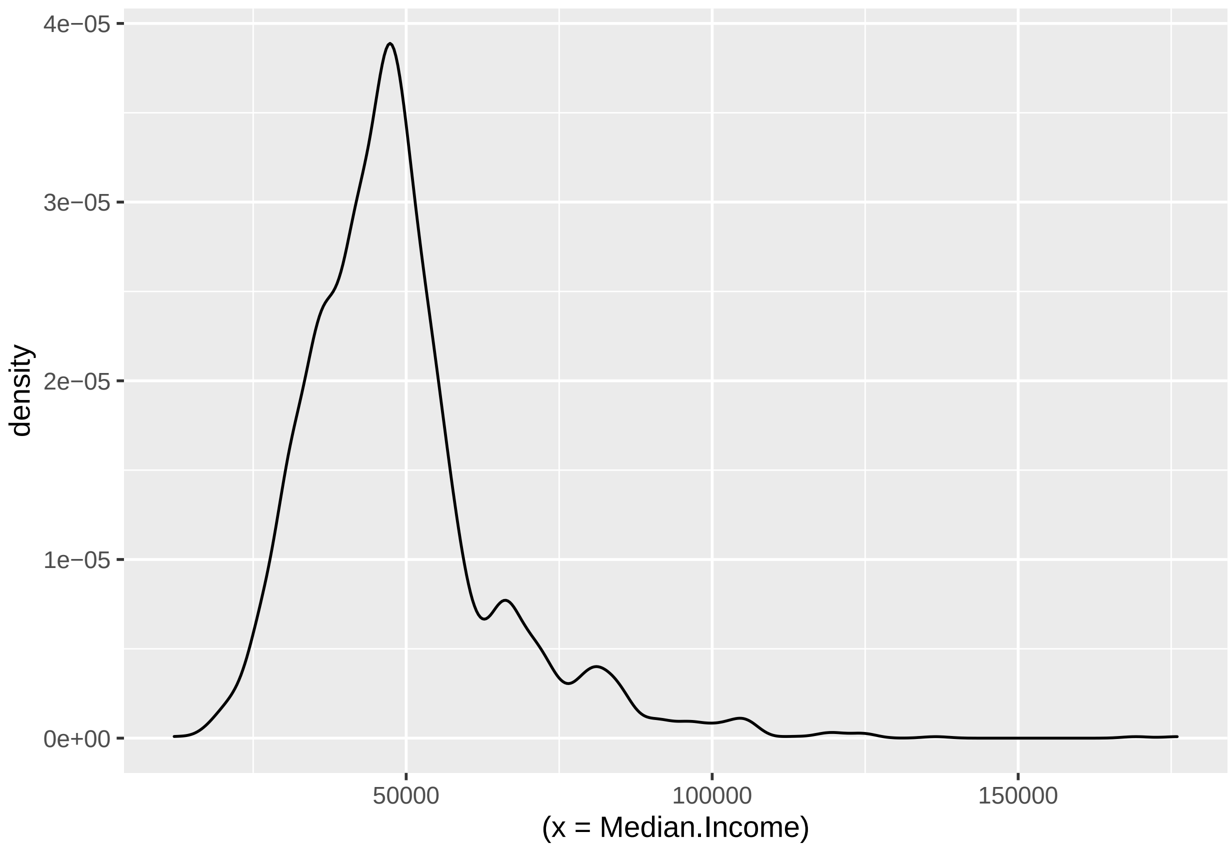
Gràfic boxplot Median.Income

```
boxplot(df_clean$Median.Income)
```



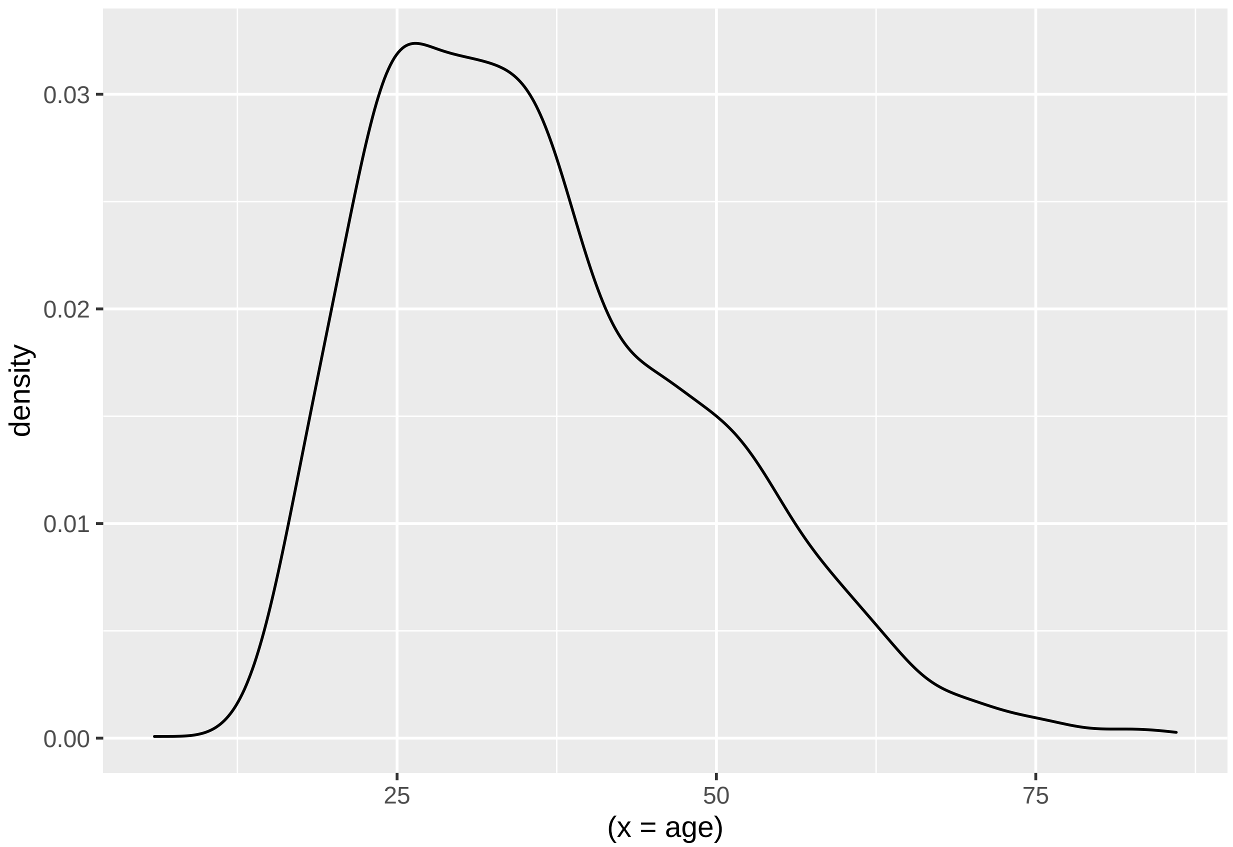
Gràfics de densitat:

```
ggplot(df_clean) + geom_density(map = aes((x = Median.Income)))
```



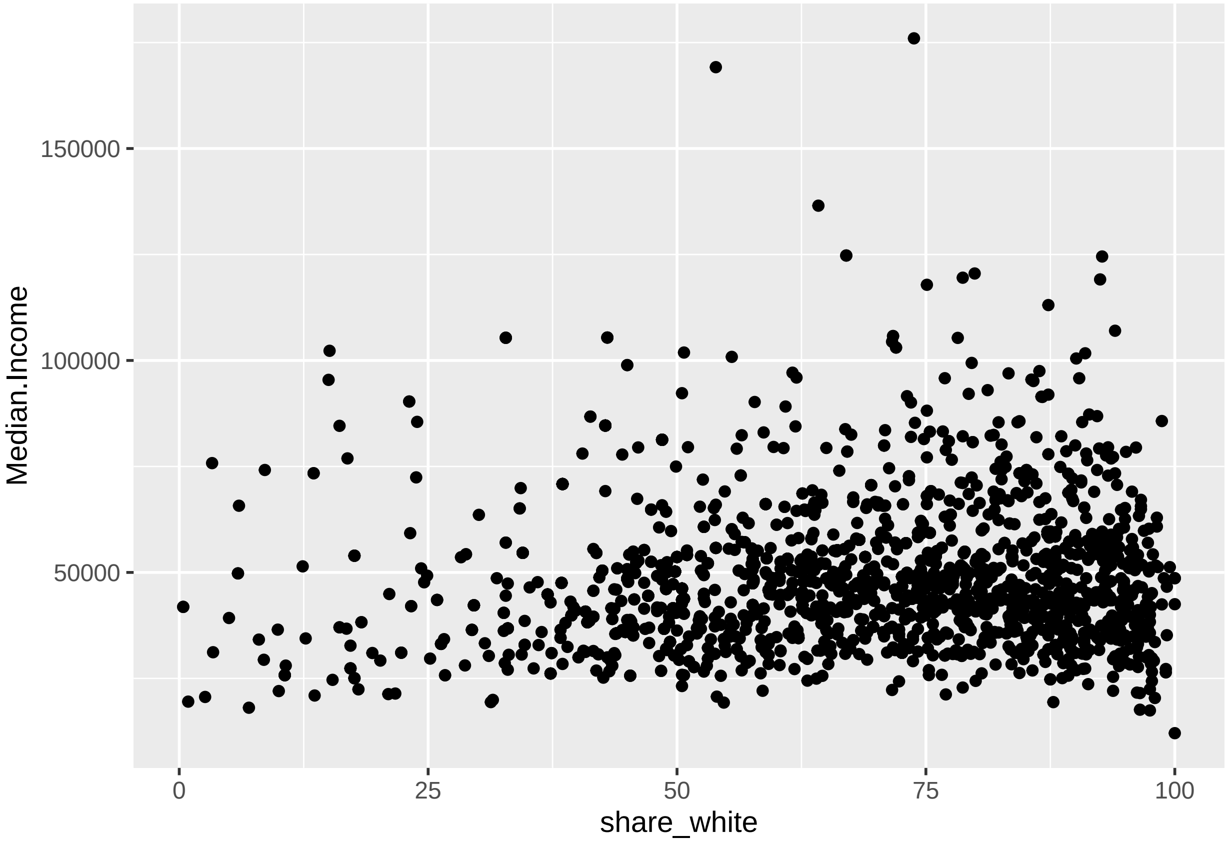
```
ggplot(df_clean) + geom_density(map = aes((x = age)))
```

```
## Warning: Removed 71 rows containing non-finite values (stat_density).
```



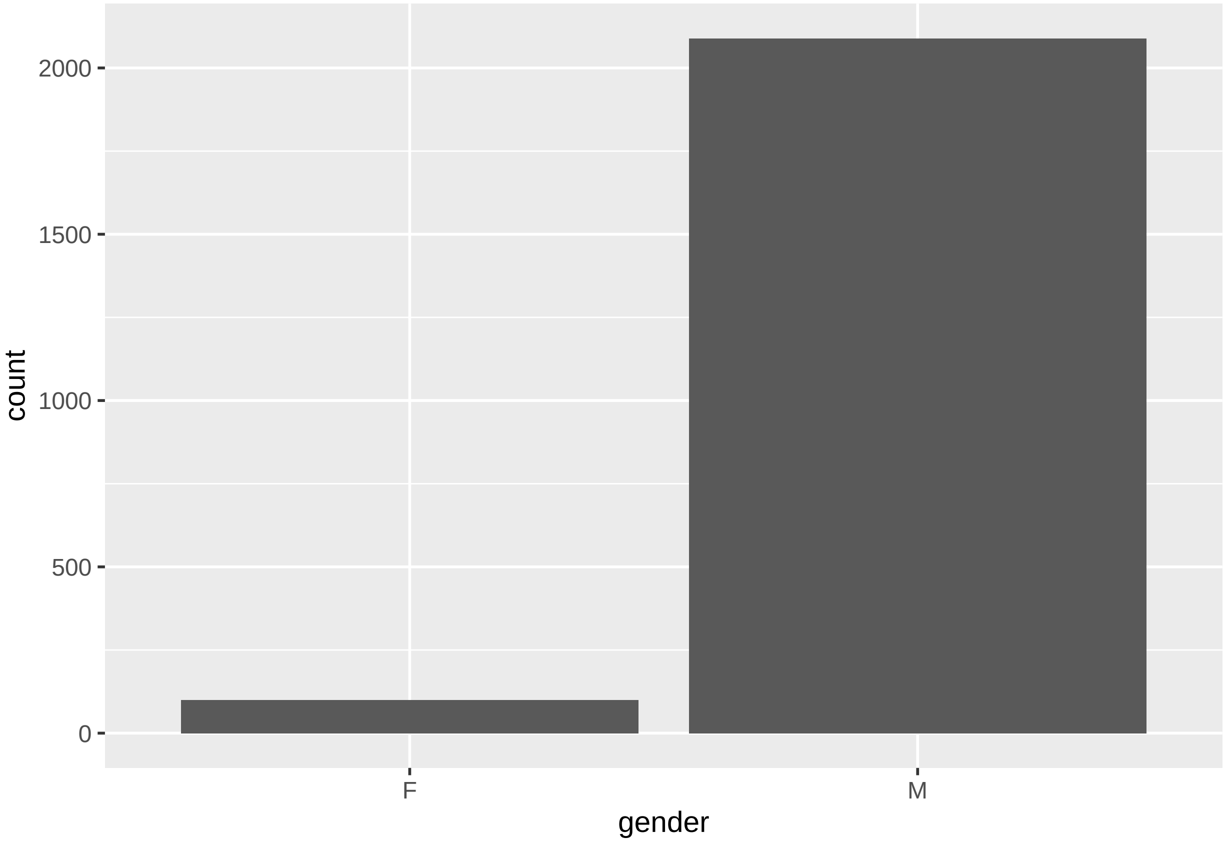
Gràfics de punts:

```
ggplot(df_clean) + geom_point(map = aes(x = share_white, y = Median.Income))
```

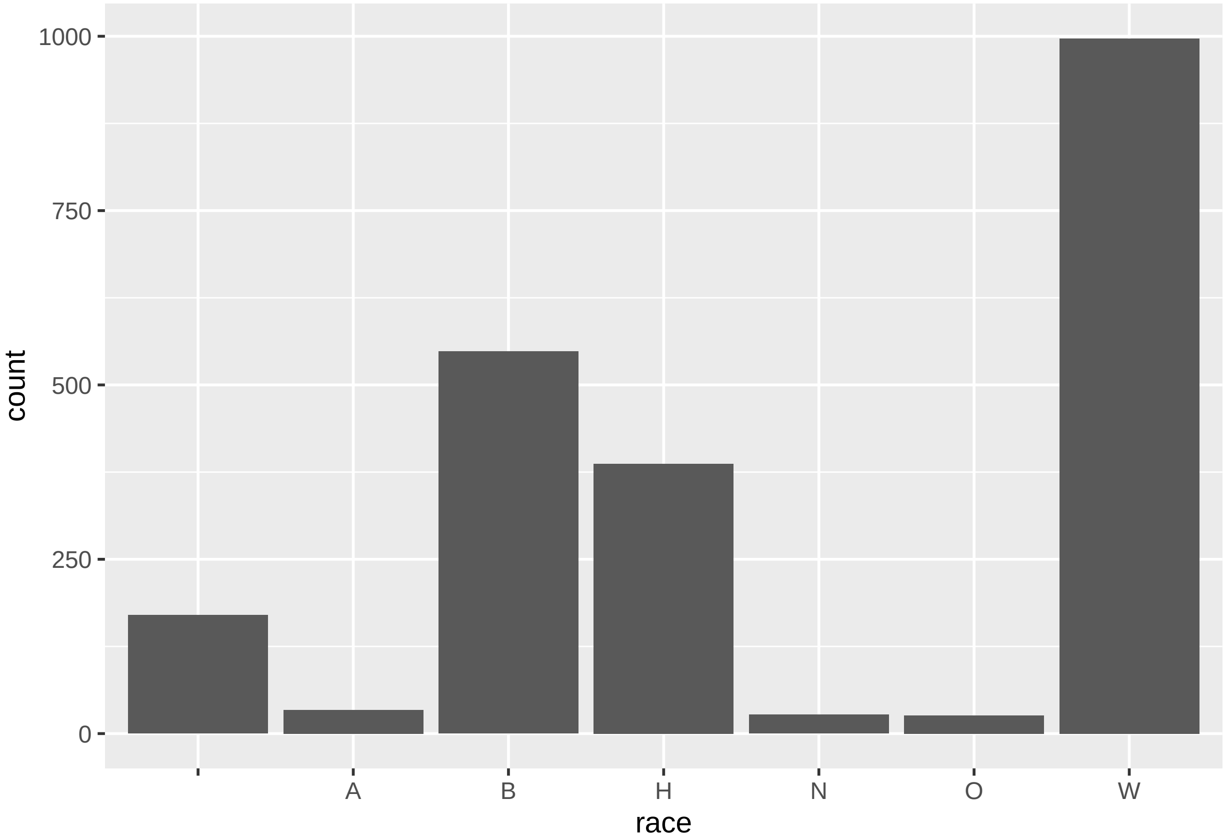


Gàfics de barres:

```
ggplot(df_clean) + geom_bar(map = aes(gender))
```



```
ggplot(df_clean) + geom_bar(map = aes(race))
```



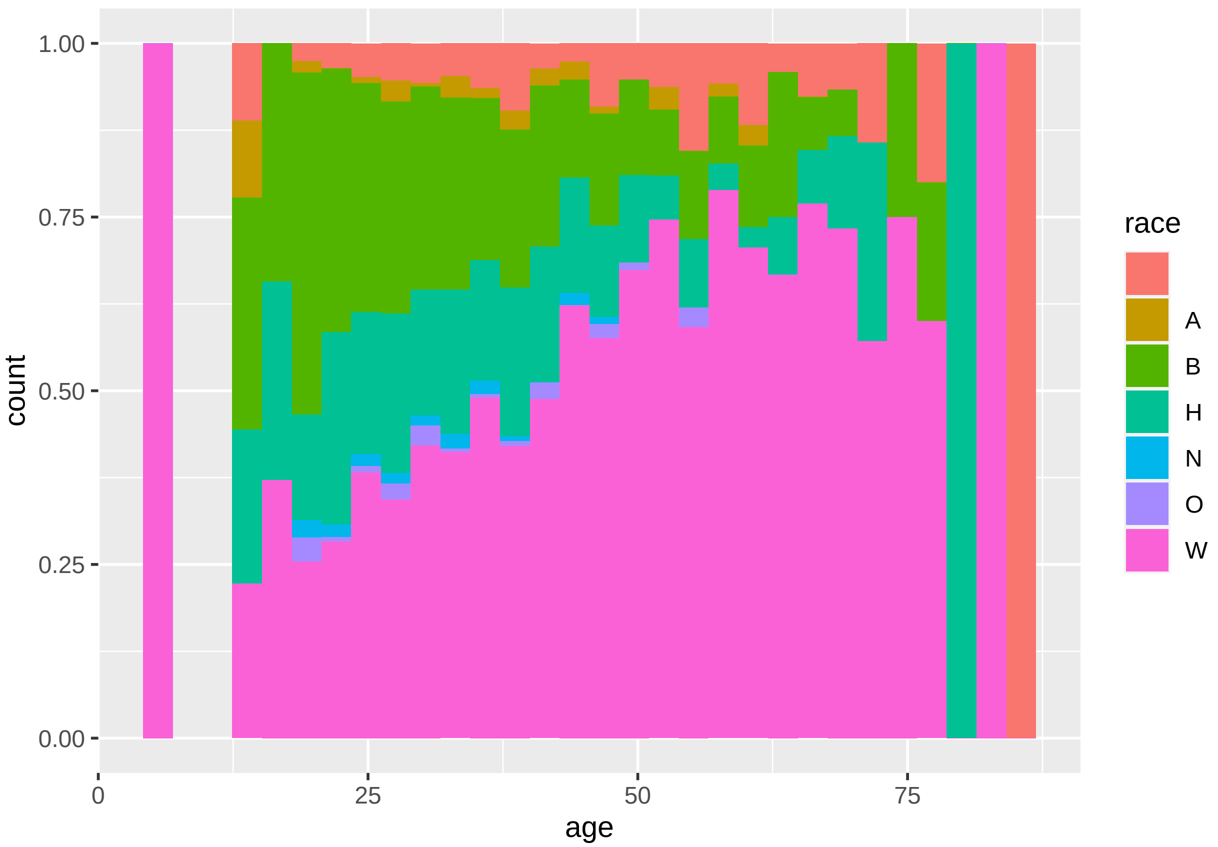
Histogrames:

```
ggplot(df_clean) + geom_histogram(map = aes(age, fill = race), position = "fill")
```

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

```
## Warning: Removed 71 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 14 rows containing missing values (geom_bar).
```

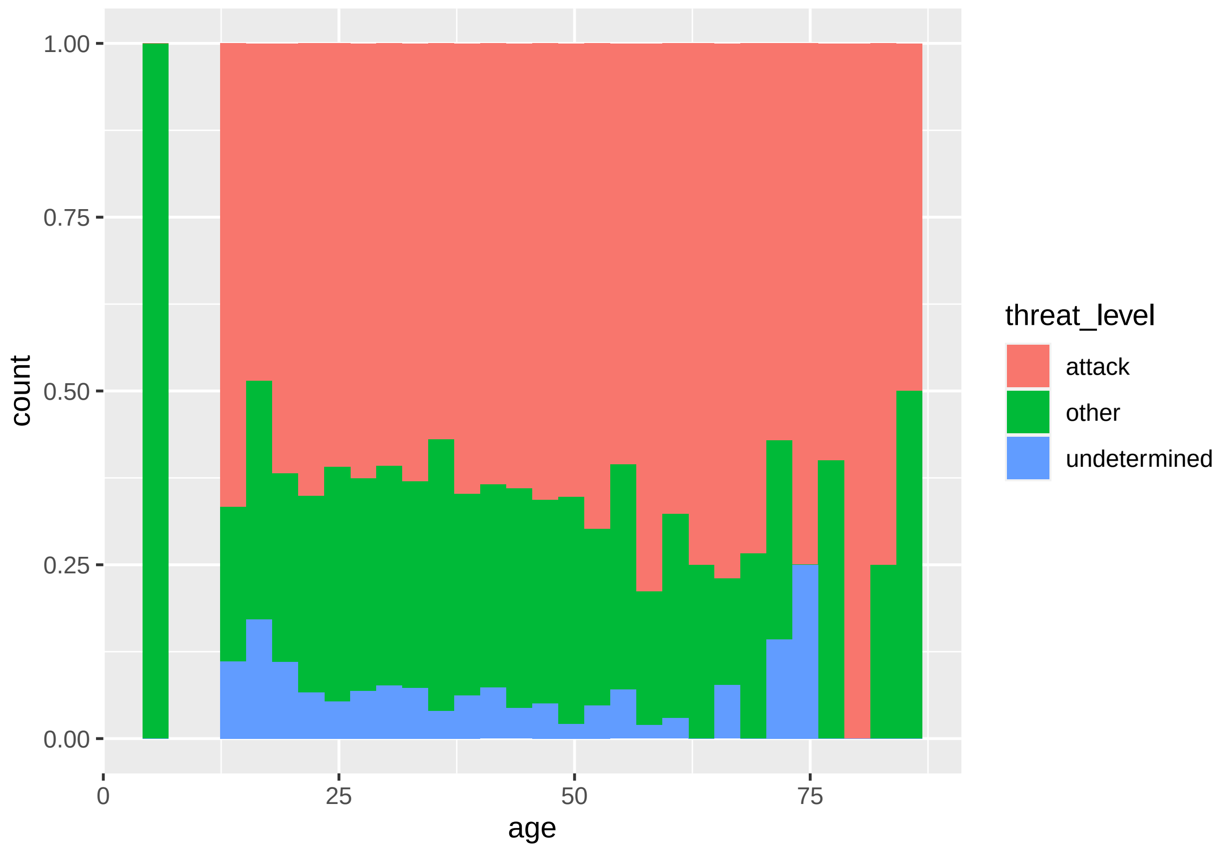


```
ggplot(df_clean) + geom_histogram(map = aes(age, fill = threat_level), position = "fill")
```

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

```
## Warning: Removed 71 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 6 rows containing missing values (geom_bar).
```

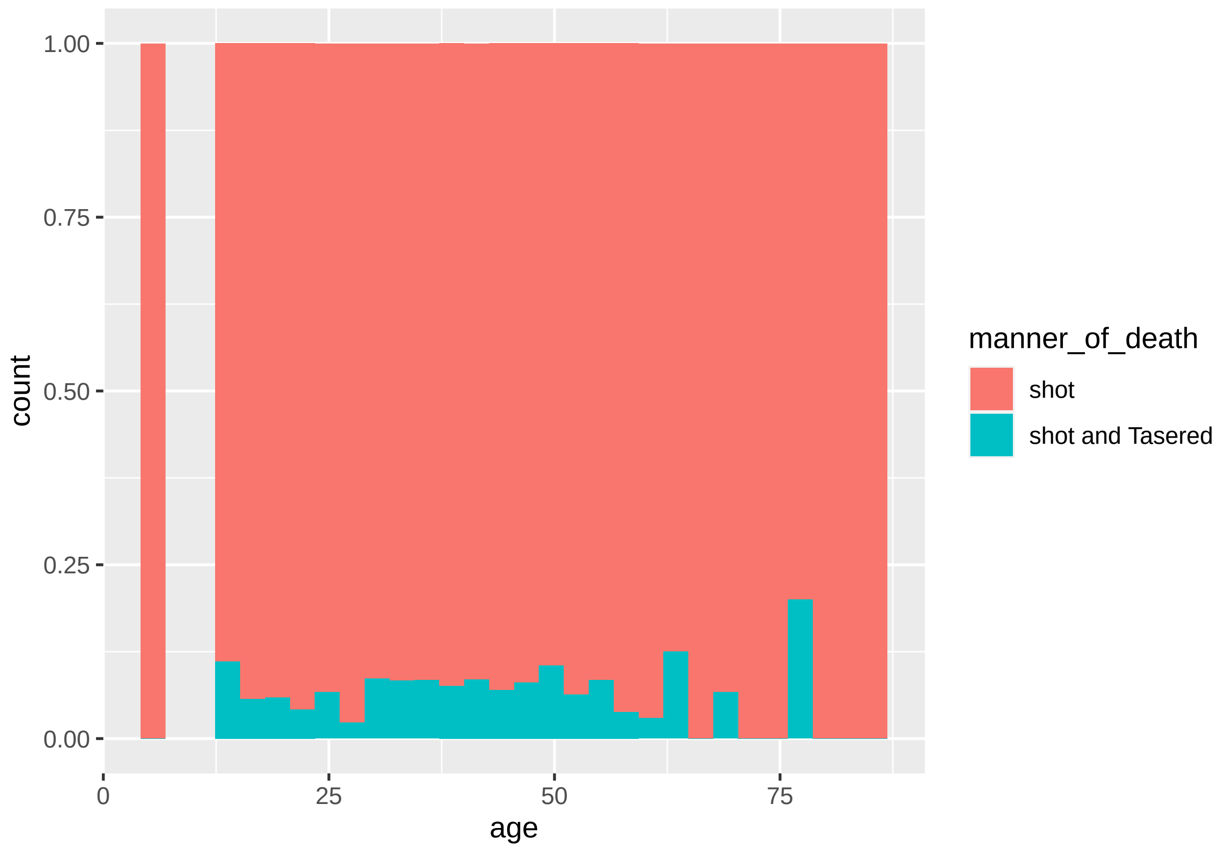


```
ggplot(df_clean) + geom_histogram(map = aes(age, fill = manner_of_death), position = "fill")
```

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

## Warning: Removed 71 rows containing non-finite values (stat_bin).

## Warning: Removed 4 rows containing missing values (geom_bar).
```


1.00
0.75
0.50
0.25
0.00
count
0
25
50
75
age
manner_of_death
shot
shot and Tasered