

Pràctica 2 - Aarón Puche i Roger Pardell

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

Aquest darrer any hem vist com han augmentat considerablement les manifestacions del moviment Black Lives Matter amb motiu de l'assassinat del ciutadà George Floyd per mans de l'actualment expolícia Derek Chauvin. Però no pareix ser un cas aïllat. Ja en 2014 hi va haver forts disturbis a la ciutat de Ferguson com a resposta per l'assassinat a mans d'una policia de Michael Brown, un jove Afroamericà de 18 anys.

Amb aquest context, el Washington Post va començar a recollir tots els tirotejos amb víctimes mortals causades per policies als Estats Units. A kaggle hem trobat un conjunt de datasets amb aquestes dades que ens poden ajudar a analitzar aquesta situació. A més de les dades dels enfrontaments amb víctimes mortals, també disposem de diversos datasets amb informació relativa a les ciutats dels Estats Units. Aquesta informació ens serà molt útil per a analitzar més en profunditat si hi ha diferències entre ciutats respecte a les ètnies predominants de cadascuna.

Algunes de les preguntes que volem respondre a partir de les dades són les següents:

- Hi ha diferències d'ingressos mitjans d'una ciutat depenent de la distribució d'ètnies.
- Hi ha ciutats on predominen persones d'unes ètnies sobre altres? Açò té alguna influència en les característiques de la ciutat?
- Les víctimes mortals pertanyen majoritàriament a alguna ètnia? En concret, hi ha diferències significatives respecte al percentatge de víctimes de població afro americana respecte a les altres?
- Predomina algun rang d'edat entre les persones que han sigut abatudes? Hi ha diferències entre ètnies?

2. Integració i selecció de les dades d'interès a analitzar.

Carreguem les llibreries que utilitzarem

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

```
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
```

```
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
```

```
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
```

```
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
```

```
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

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

3. Neteja de les dades.

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 substituïm 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 assignem als valors que havíem substituït 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)

head(df_clean)
```

```
##           name      date manner_of_death armed age gender race
## 1   Vincent Nageak 2016-02-10          shot   gun  36      M    N
## 2   Jean R. Valescot 2017-02-17          shot   gun  35      M    B
## 3 Matthew Colton Stover 2017-06-19          shot   gun  21      M    N
## 4   Tristan Vent 2015-09-08          shot   gun  19      M    N
## 5   Vincent J. Perdue 2015-09-09          shot   gun  33      M    N
## 6 James Robert Richards 2016-08-29 shot and Tasered   gun  28      M
##  signs_of_mental_illness threat_level      flee body_camera
## 1                FALSE      attack Not fleeing      FALSE
## 2                FALSE      attack Not fleeing      FALSE
## 3                 TRUE      attack      Foot      FALSE
## 4                FALSE      attack Not fleeing      FALSE
## 5                FALSE      attack      Car      FALSE
## 6                FALSE      attack      Foot      TRUE
##  percent_completed_hs poverty_rate Median.Income share_white share_black
## 1                84.6         11.7       76902         16.9         1.0
## 2                90.4          9.6       70988         86.1         0.2
## 3                91.2         13.1       55229         66.1         9.0
## 4                91.2         13.1       55229         66.1         9.0
## 5                91.2         13.1       55229         66.1         9.0
## 6                91.2         13.1       55229         66.1         9.0
##  share_native_american share_asian share_hispanic
## 1                61.2          9.1          3.1
## 2                 7.0          0.5          3.1
## 3                10.0          3.6          9.0
## 4                10.0          3.6          9.0
## 5                10.0          3.6          9.0
## 6                10.0          3.6          9.0
```

Valors buits i extrems

Una vegada tenim les dades en el format que volem, comprovem si hi ha algun valor buit entot el dataframe:

```
colSums(is.na(df_clean))
```

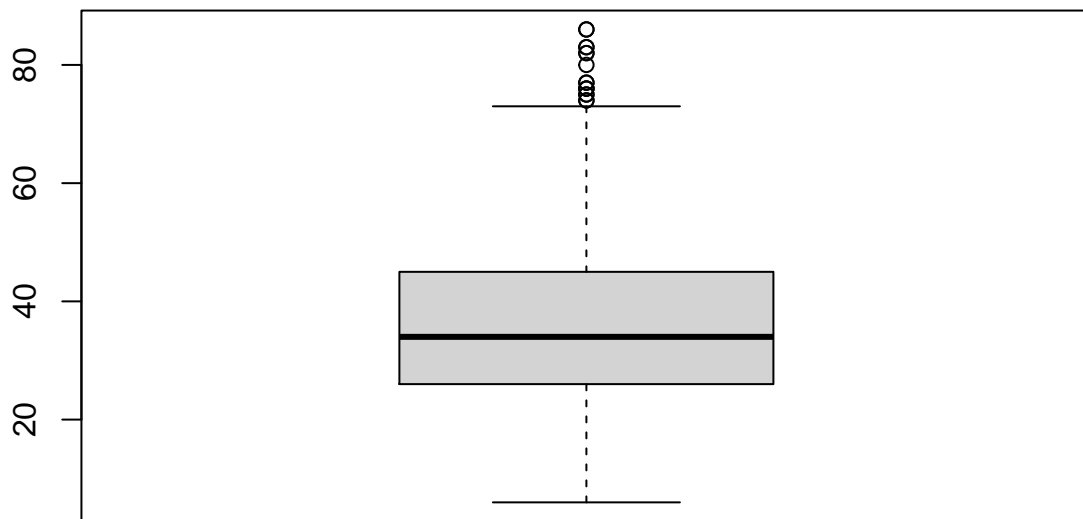
```
##           name      date      manner_of_death
##           0           0              0
##      armed           age           gender
##           0           71              0
##      race signs_of_mental_illness      threat_level
##           0              0              0
##      flee      body_camera percent_completed_hs
##           0              0              0
##      poverty_rate      Median.Income      share_white
##           0              0              0
##      share_black share_native_american      share_asian
##           0              0              0
##      share_hispanic
```

```
##          0
```

age

Vegem que tenim 71 edats desconegudes.

```
boxplot(df_clean$age)$out
```



```
## [1] 77 83 86 82 74 75 75 76 76 83 77 80 86 74 76 82
```

Tenim uns quants valors extrems en la variable edat, que corresponen a edats majors de 70 anys. Com aquest valors podrien influir en la mitjana, anema a assignar als valors NA la mediana, que es mes robusta contra aquests efectes.

```
df_clean$age[is.na(df_clean$age)] <- median(df_clean$age[!is.na(df_clean$age)])  
colSums(is.na(df_clean))
```

```
##          name          date    manner_of_death  
##          0            0            0  
##        armed          age      gender  
##          0            0            0  
##        race signs_of_mental_illness  threat_level  
##          0            0            0  
##        flee      body_camera  percent_completed_hs
```

```
##           0           0           0
##      poverty_rate      Median.Income      share_white
##           0           0           0
##      share_black  share_native_american      share_asian
##           0           0           0
##      share_hispanic
##           0
```

Ya no tenim valors NA en cap variable. Comprovem si hi ha elements amb cadena buida.

```
colSums(df_clean == "")
```

```
##           name           date      manner_of_death
##           0           NA           0
##      armed           age           gender
##           8           0           0
##      race signs_of_mental_illness      threat_level
##      170           0           0
##      flee      body_camera      percent_completed_hs
##      54           0           0
##      poverty_rate      Median.Income      share_white
##           0           0           0
##      share_black  share_native_american      share_asian
##           0           0           0
##      share_hispanic
##           0
```

date

Ens indica que hi ha valors NA en date, ho comprovem.

```
which(is.na(df_clean$date))
```

```
## integer(0)
```

En la comprovació ens diu que no hi ha cap valor de date amb NA.

armed

Com són sols 8 registres, els eliminarem

```
df_clean <- df_clean[df_clean$armed != "",]
```

race

Aquestes dades serà una mica més complicat, ja que no són valors numèrics, si no classes a les que pot pertanyer el registre. Vegem com es distribueixen.

```
table(df_clean$race)
```

```
##
##      A  B  H  N  O  W
## 167 34 545 386 27 26 996
```

Qui mes tenim es de gent “White” pero no podem assignar els 167 registres a white ja que aço podria després fer-nos caure en analisis erronis ja que estariem inflant aquestes dades. Com no es una gran quantitat de registres respecte al total, els eliminarem.

```
df_clean <- df_clean[df_clean$race != "",]
```

flee

```
table(df_clean$flee)
```

```
##
##           Car           Foot Not fleeing           Other
##           44           319           249           1324           78
```

Tenim una situació semblant a la variable race. Com son sols 44 registres els eliminarem.

```
df_clean <- df_clean[df_clean$flee != "",]
```

En aquest punt ja tenim les dades tractades. Comprovem:

```
colSums(df_clean == "")
```

```
##           name           date           manner_of_death
##           0           NA           0
##           armed           age           gender
##           0           0           0
##           race signs_of_mental_illness           threat_level
##           0           0           0
##           flee           body_camera           percent_completed_hs
##           0           0           0
##           poverty_rate           Median.Income           share_white
##           0           0           0
##           share_black share_native_american           share_asian
##           0           0           0
##           share_hispanic
##           0
```

No tenim cap valor buit.

```
str(df_clean)
```

```
## 'data.frame':   1970 obs. of  19 variables:
## $ name           : chr  "Vincent Nageak" "Jean R. Valescot" "Matthew Colton Stover" "Tristan
## $ date           : Date, format: "2016-02-10" "2017-02-17" ...
## $ manner_of_death : Factor w/ 2 levels "shot","shot and Tasered": 1 1 1 1 1 1 1 1 1 1 ...
## $ armed          : Factor w/ 62 levels "", "air conditioner",...: 20 20 20 20 20 20 20 20 27 2
## $ age            : num  36 35 21 19 33 23 38 36 33 29 ...
## $ gender         : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
## $ race           : Factor w/ 7 levels "", "A", "B", "H",...: 5 3 5 5 5 7 5 7 7 3 ...
## $ signs_of_mental_illness: logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
## $ threat_level    : Factor w/ 3 levels "attack","other",...: 1 1 1 1 1 1 1 1 2 1 ...
```



```
## $ flee          : Factor w/ 5 levels "", "Car", "Foot", ...: 4 4 3 4 2 2 4 5 4 4 ...
## $ body_camera   : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ percent_completed_hs : num 84.6 90.4 91.2 91.2 91.2 91.2 90.2 91.8 91.8 69.1 ...
## $ poverty_rate   : num 11.7 9.6 13.1 13.1 13.1 13.1 14.8 11.7 11.7 29.1 ...
## $ Median.Income  : num 76902 70988 55229 55229 55229 ...
## $ share_white     : num 16.9 86.1 66.1 66.1 66.1 66.1 82.2 83.4 83.4 54.4 ...
## $ share_black     : num 1 0.2 9 9 9 9 0.4 1.4 1.4 41.4 ...
## $ share_native_american : num 61.2 7 10 10 10 10 6.7 5.2 5.2 0.1 ...
## $ share_asian     : num 9.1 0.5 3.6 3.6 3.6 3.6 0.6 2.1 2.1 1 ...
## $ share_hispanic  : num 3.1 3.1 9 9 9 9 3.3 4.3 4.3 3.1 ...
```

I finalment, comprovem que estan amb el tipus de dades adequat adequat.

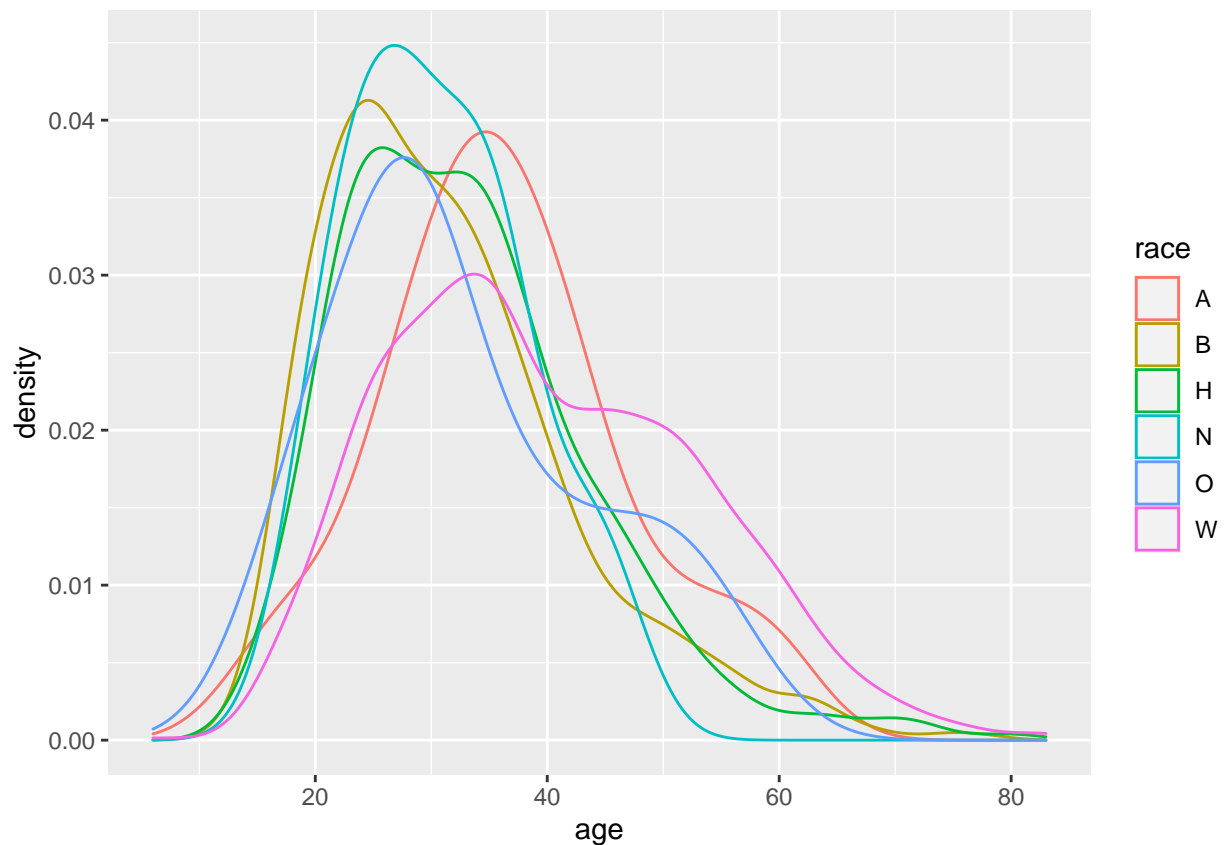
4 i 5. Anàlisi de les dades i representació dels resultats a partir de taules i gràfiques.

Analisis exploratori de les dades

En aquest punt ens centrarem en les variables gender, race, poverty, hig school i age.

En primer lloc mirarem si la distribució de l'edat es igual a totes les races:

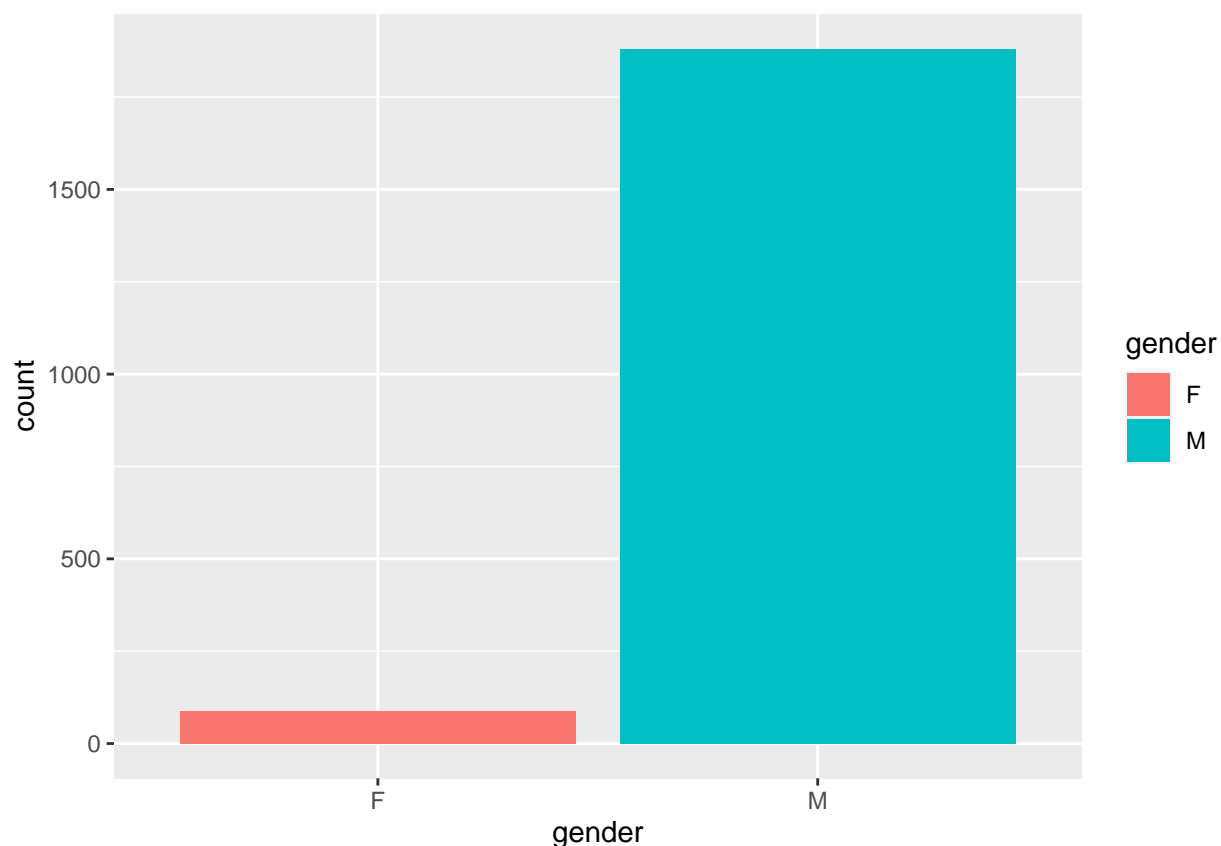
```
#ggplot(df_clean) + geom_density(map = aes((x = age)))
ggplot(df_clean, aes(x = age, colour=race, group=race)) + geom_density()
```



Com es pot observar, les persones disparades de raça Black, Hispanic i Native-Americans tenen una distribució bastant similars d'edats. Assian es una mica més majors però on es veu una diferència més gran és amb White, aquestes últimes persones són més majors, no s'agrupen tant en dades entorn als 20-30 anys com els altres.

Vegem també la distribució per sexes.

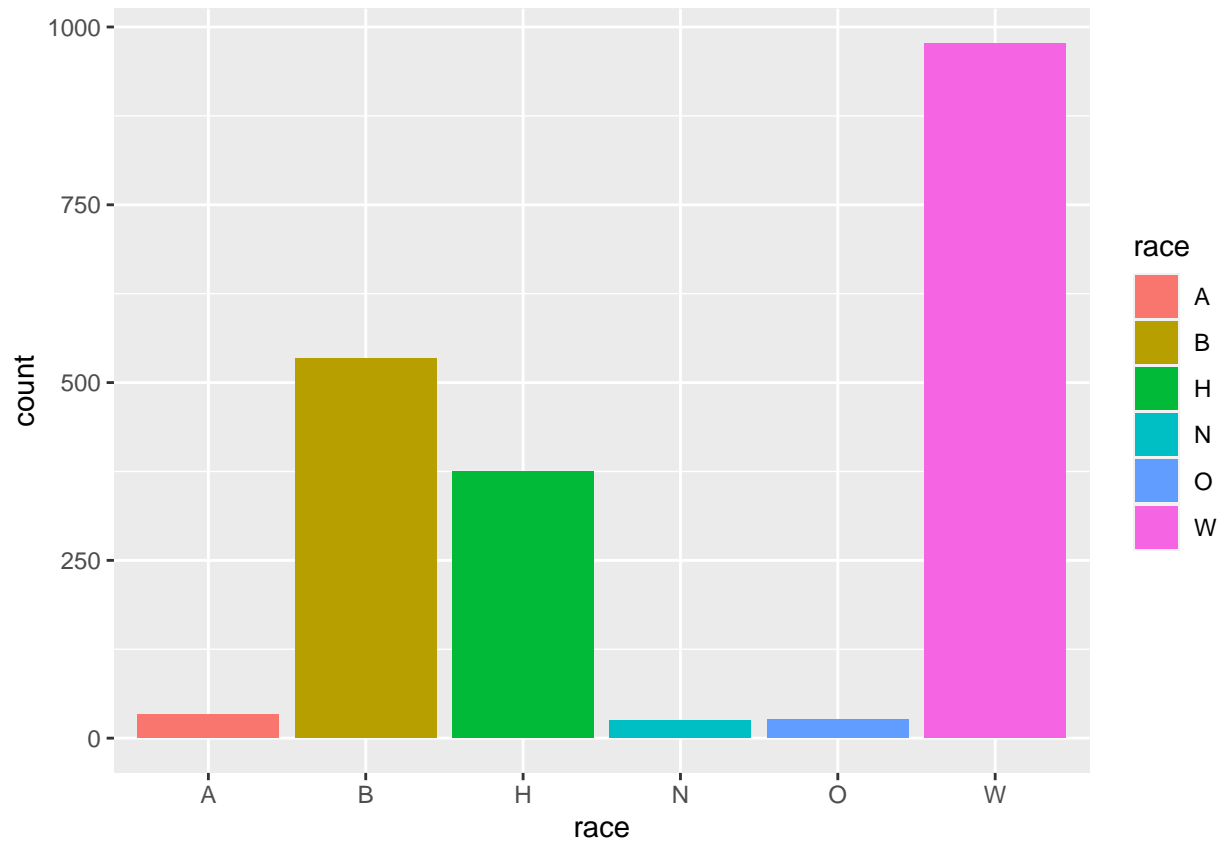
```
ggplot(df_clean) + geom_bar(mapping = aes(gender, fill=gender))
```



Podem destacar que, clarament, la gran majoria de persones disparades són homes.

Vegem ara com es distribueixen les races de les persones disparades:

```
ggplot(df_clean) + geom_bar(mapping = aes(race, fill=race))
```



De qui mes registres tenim es de raça “blanca”. Es llogic veure aquests resultats ja que la majoria de persones en estats units son de raça blanca, però, quina es la proporció de víctimes de cada respectiva raça respecta al total de persones d'eixa raça?

Per saber el total de la població de cada raça ens basarem en les dades de Demographics of the United States, de 2017. Les nostres dades, com podem veure, van de 2015 a 2017, així que s'ajustaran bastant be.

```
summary(df_clean$date)
```

```
##           Min.         1st Qu.         Median         Mean         3rd Qu.         Max.
## "2015-01-02" "2015-08-06" "2016-03-10" "2016-03-20" "2016-11-03" "2017-07-31"
```

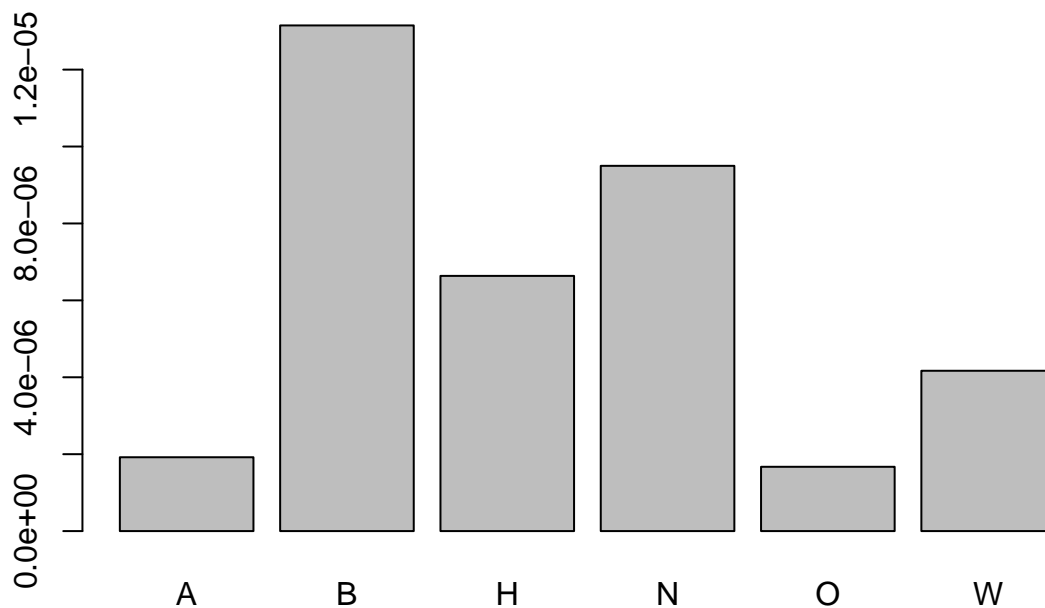
La data mínima de les dades es 02/01/2015 i la màxima, 31/07/2017

Vegem el gràfic de les víctimes de disparats tenint en compte la població total:

```

races_total <- c((sum(df_clean$race == 'A') / 17186320), (sum(df_clean$race == 'B') / 40610815), (sum(d
races <- c('A', 'B', 'H', 'N', 'O', 'W')

barplot(races_total, names.arg = races)
```



Vegent el grafic amb les dades comparades amb el total de poblacio de cada raça pareix indicar que les persones de raça negra, nadius americans i hispanos son mes probables de rebre un tir d'un policia que una persona blanca.

Per analitzar un poc mes en profunditat algunes posibles relacions entre les persones que han sigut disparadaes, discretizarem les variables `percent_completed_hs` i `poverty_rate`. Dividirem els seus valors en les franjes: 0-24, 25-49, 50-74 i 75-100.

```
df_clean["range_highSchool"]<- cut(df_clean$percent_completed_hs, breaks = c(0,24,49,74,100), labels = c("0 - 24", "25 - 49", "50 - 74", "75 - 100"))
df_clean["range_poverty"]<- cut(df_clean$poverty_rate, breaks = c(0,24,49,74,100), labels = c("0 - 24", "25 - 49", "50 - 74", "75 - 100"))
```

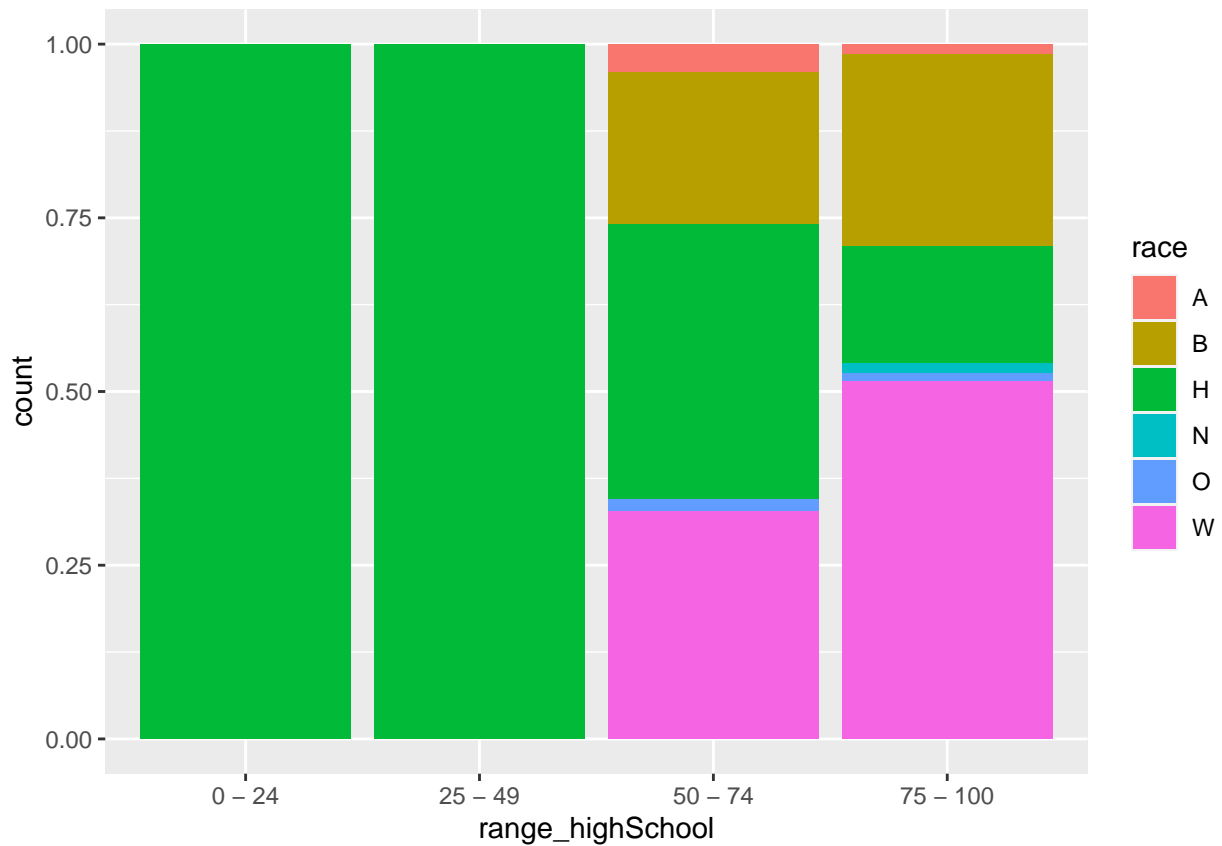
```
table(df_clean$range_highSchool)
```

```
##
##  0 - 24  25 - 49  50 - 74  75 - 100
##      1       6     174    1789
```

```
table(df_clean$range_poverty)
```

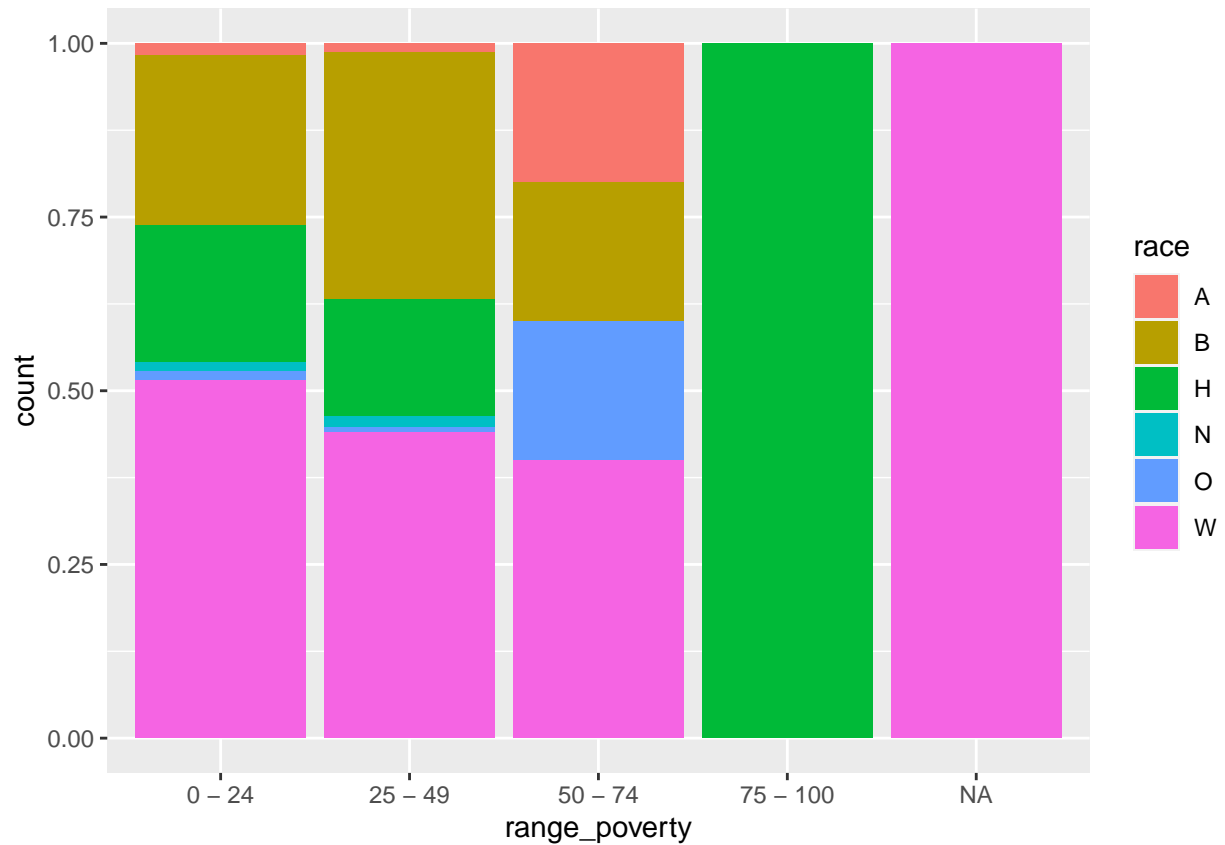
```
##
##  0 - 24  25 - 49  50 - 74  75 - 100
##  1479    484      5       1
```

```
ggplot(df_clean) + geom_bar(map = aes(x = range_highSchool, fill=race), position = "fill")
```



Ens fixem amb els percentatges entre 50 i 100% perquè en els anteriors hi han molt poques dades. Pareix que les víctimes blanques solen viure en ciutats amb un nivell d'estudis superiors. Destaca sobretot la diferència amb la gent hispana.

```
ggplot(df_clean) + geom_bar(map = aes(x = range_poverty, fill=race), position = "fill")
```



Per ultim, mitjançant histogrames vegem com es distribueixen els tipus d'amenaça que s'indiquin i la manera en que ha mort la persona:

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

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

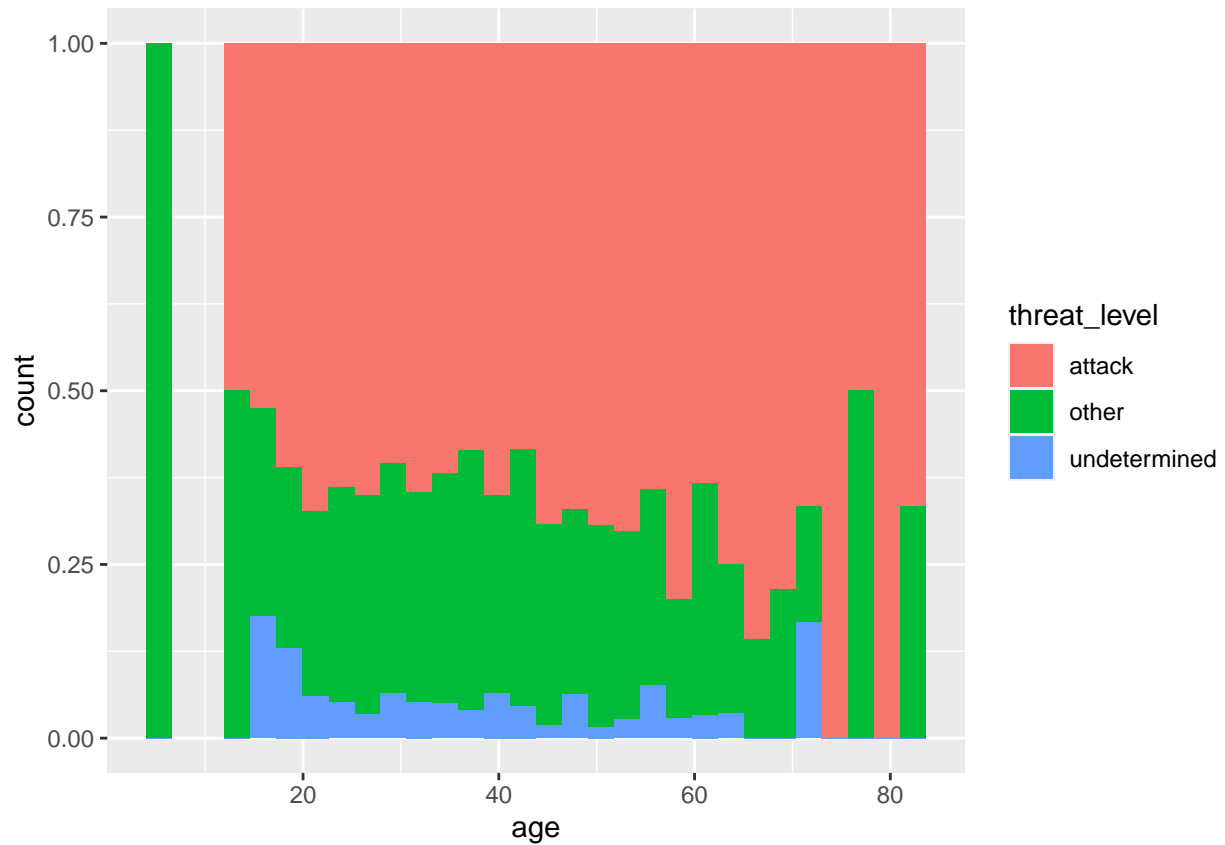
```
## Warning: Removed 12 rows containing missing values (geom_bar).
```



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

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

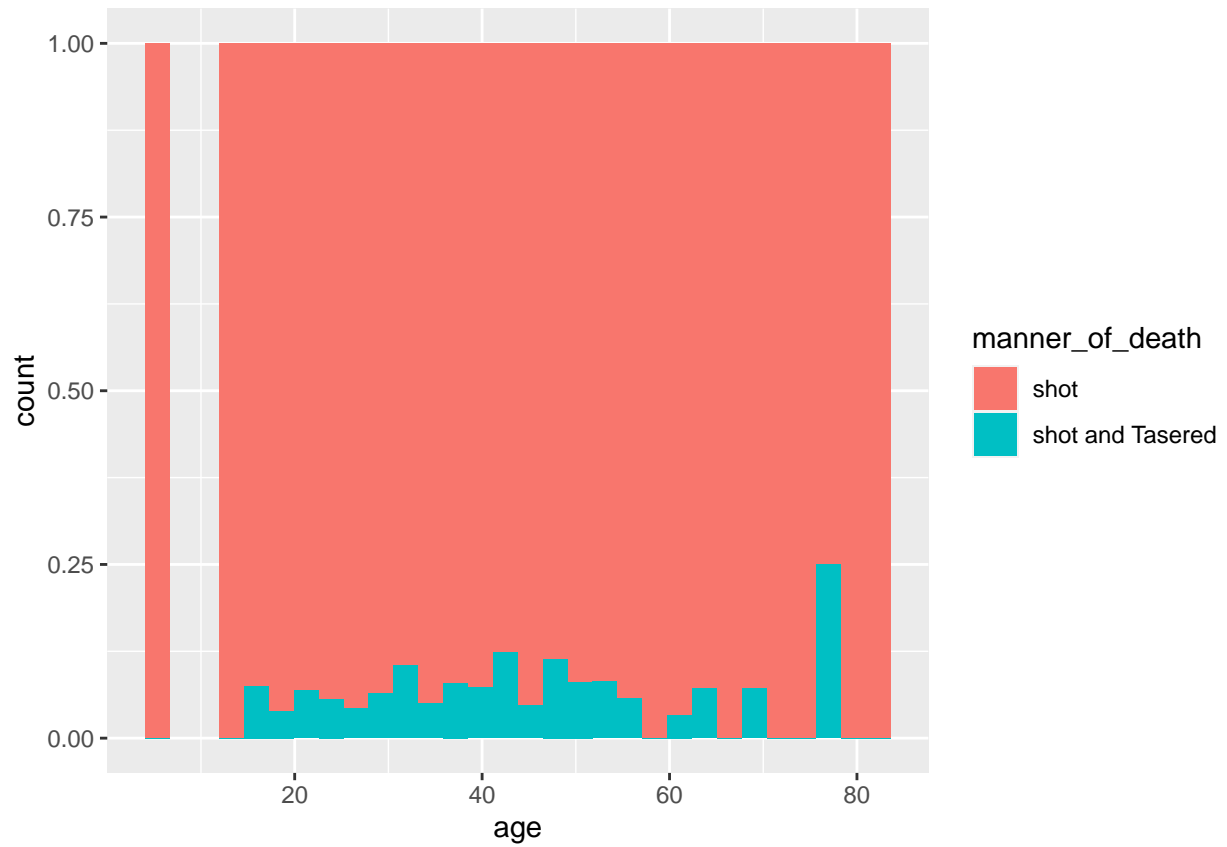
```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



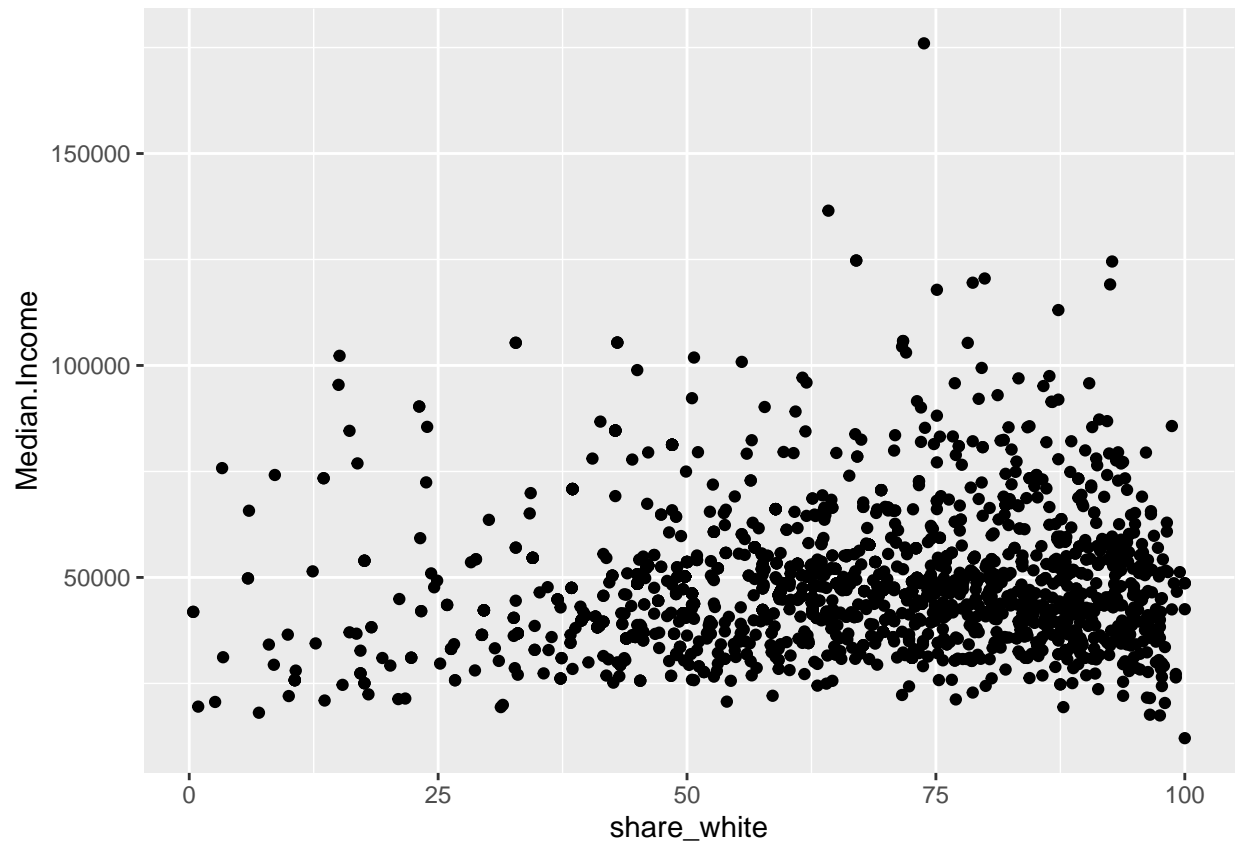
```
ggplot(df_clean) + geom_histogram(map = aes(age, fill = manner_of_death), position = "fill")
```

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

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

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



6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?