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 en mans de l'actualment expolicia Derek Chauvin. Però no pareix ser un cas aïllat. Ja durant el 2014 hi va haver forts disturbis a la ciutat de Fergusson com a resposta per l'assassinat a mans d'una policia de Michael Brown, un jove Afroamericà de 18 anys.

En 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 quatre datasets més amb informació relativa a les ciutats dels Estats Units. Aquesta informació ens serà molt útil per a analitzar més en profunditat si hi ha diferencies 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?
- Predomina algun rang d'edat entre les persones que han sigut abatudes? Hi ha diferències entre ètnies?

Així, els datasets d'interés són els següents:

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

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
```

```
dim(df_policeKilling)
```

```
## [1] 2535   14
```

```
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
```

```
dim(df_householdIncome)
```

```
## [1] 29322     3
```

```
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
```

```
dim(df_poverty)
```

```
## [1] 29329     3
```

```
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
```

```
dim(df_highSchool)
```

```
## [1] 29329     3
```

```
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
```

```
dim(df_shareRace)
```

```
## [1] 29268     7
```

El primer mostra totes les víctimes mortals en mans de la policia. En total, tenim 2535 observacions i 14 atributs, entre els quals hi trobem l'edat, l'estat i la ciutat on ha succeït i la manera en què han mort.

Els altres quatre conjunts són formats per unes 29300 observacions (entre 29268 i 29329), i un total de 10 atributs diferents. Cada observació és una ciutat americana, i els atributs són valors demogràfics i econòmics d'aquestes ciutats, per exemple, el percentatge de persones segons raça, la mediana dels ingressos familiars, etc.

Aquests cinc conjunts aporten les característiques individuals de les víctimes, així com les variables ambientals de les ciutats on han succeït els assassinats. Així, se'ns permet fer una anàlisi més holística de la situació.

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

Per a fer les nostres anàlisis crearem un dataset un amb totes les dades referents a informació sobre les ciutats relacionades amb la informació de les persones assassinades.

En primer lloc carreguem les llibreries que utilitzarem al llarg de la pràctica:

```
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(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(magrittr)
```

Canviem el nom de la columna Geographic.area per a després fer la mescla de les dades:

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

Mesclem els distints datasets i obtenim el dataset USA que contindrà tota la informació respectiva a les ciutats:

```
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 per fer la mescla de la informació de les ciutats i les persones assassinades:

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

Mesclem del dataset obtingut amb df_policeKilling:

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

# Eliminem aquesta, ja que en aquest cas no ens aporta informacio rellevant
df_clean$id <- 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 substituim 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)
head(df_clean)
```

```
##   state      city                  name       date  manner_of_death armed age
## 1    AK    Barrow        Vincent Nageak 2016-02-10             shot   gun  36
## 2    AK   BigLake      Jean R. Valescot 2017-02-17             shot   gun  35
## 3    AK Fairbanks Matthew Colton Stover 2017-06-19             shot   gun  21
## 4    AK Fairbanks          Tristan Vent 2015-09-08             shot   gun  19
## 5    AK Fairbanks     Vincent J. Perdue 2015-09-09             shot   gun  33
## 6    AK Fairbanks James Robert Richards 2016-08-29 shot and Tasered   gun  28
##   gender race signs_of_mental_illness threat_level        flee body_camera
## 1      M    N                  FALSE       attack Not fleeing       FALSE
## 2      M    B                  FALSE       attack Not fleeing       FALSE
## 3      M    N                   TRUE       attack        Foot       FALSE
## 4      M    N                  FALSE       attack Not fleeing       FALSE
## 5      M    N                  FALSE       attack         Car       FALSE
## 6      M                       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:

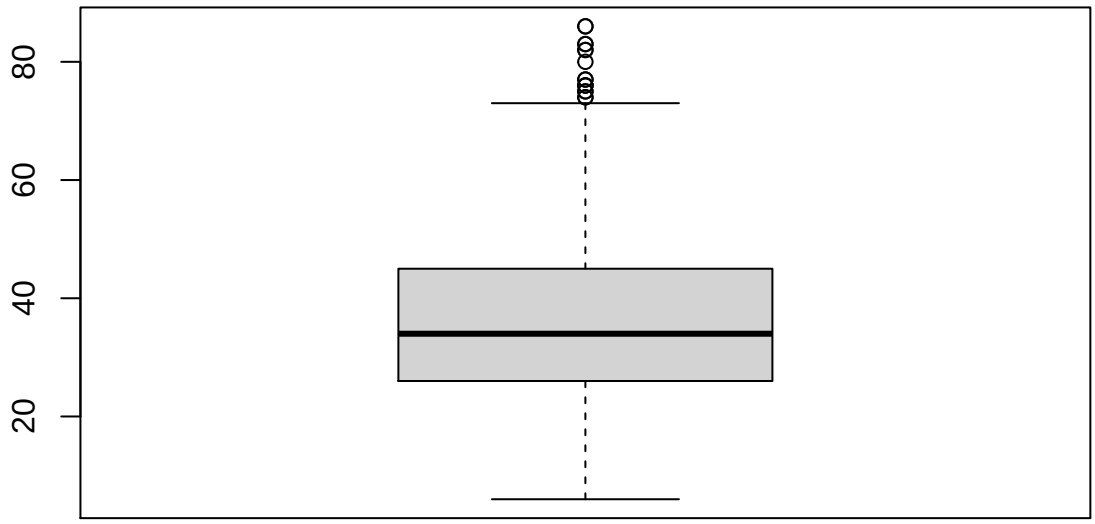
```
colSums(is.na(df_clean))
```

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

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

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

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

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

armed

Com son sols 8 registres, els eliminarem

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

race

Aquestes dades serà una mica mes complicat, ja que no son 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 despres 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 == "")
```

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

No tenim cap valor buit.

```
str(df_clean)
```

```
## 'data.frame':    1970 obs. of  21 variables:
##  $ state                   : chr  "AK" "AK" "AK" "AK" ...
##  $ city                    : chr  "Barrow" "BigLake" "Fairbanks" "Fairbanks" ...
##  $ name                    : chr  "Vincent Nageak" "Jean R. Valescot" "Matthew Colton Stover" "Tristar
##  $ 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 1
##  $ 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 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.

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

En aquest punt hem decidit ajuntar els punts 4 i 5 de l'enunciat per a fer en primer lloc una anàlisi mitjançant taules i gràfiques de les relacions que podem observar d'algunes variables que ens interessin i després passar a fer anàlisis amb més profunditat de les dades.

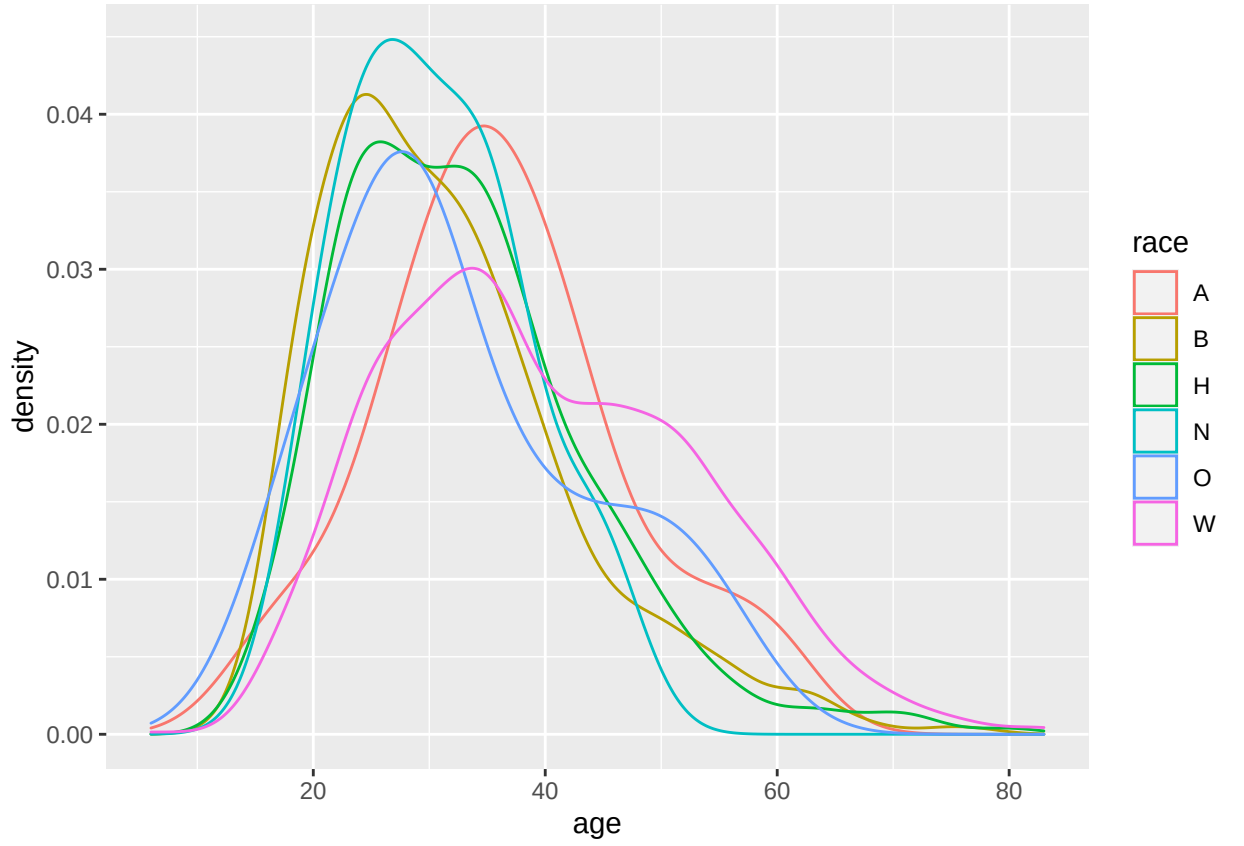
Analisis exploratori de les dades

Ens centrarem en les relacions de les següents característiques: edat, gènere, ètnia, nivell de pobresa, nivell d'estudis, ingressos mitjans, el tipus d'amenaça i la forma de morir.

Distribució per edats

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

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



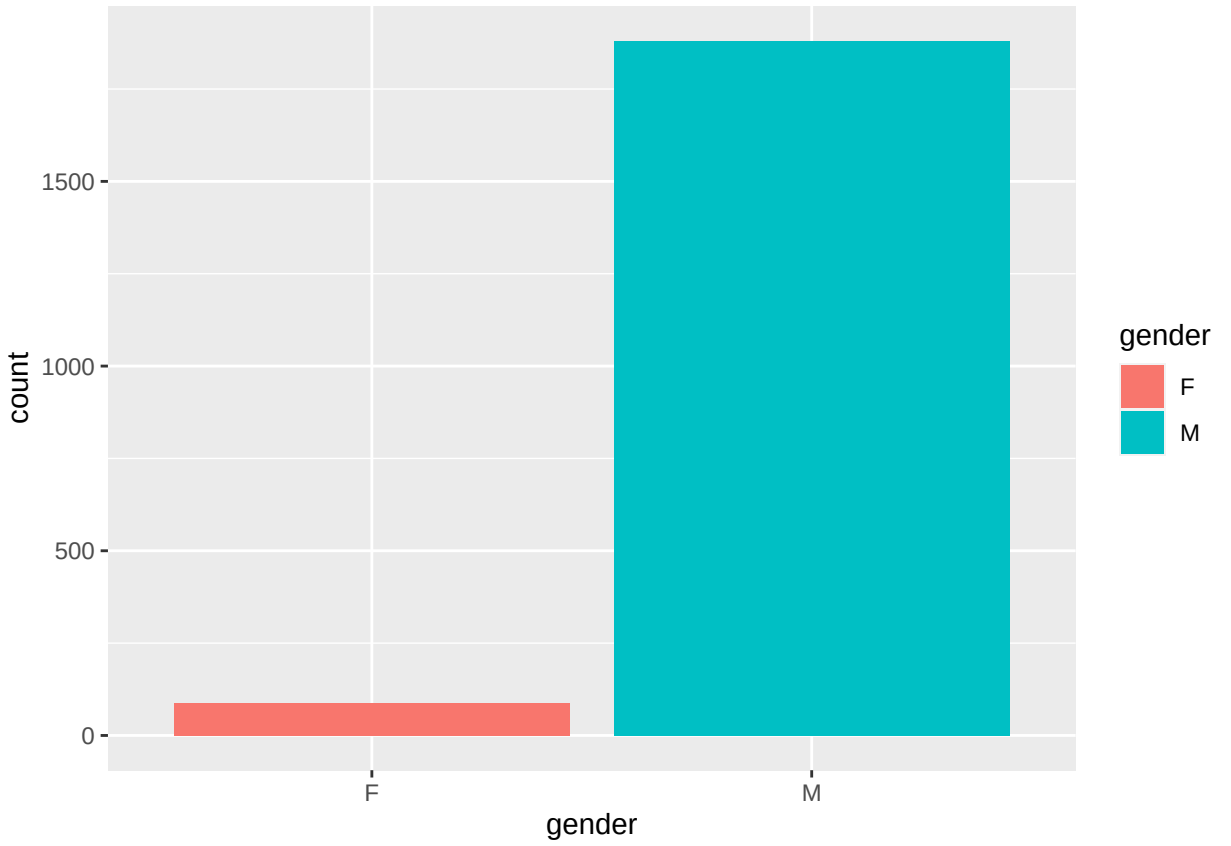
Com es pot observar, les persones assassinades d'ètnia negra, hispana o nativa americana tenen una

distribució de les edats bastant similar. Les edats de les persones asiàtiques són una mica més majors, però on es veu una diferència més gran és amb les persones blanques. Aquestes persones són més majors, no s'agrupen tant entorn als 20-30 anys com les altres.

Distribució per gèneres

Vegem també la distribució per sexes:

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

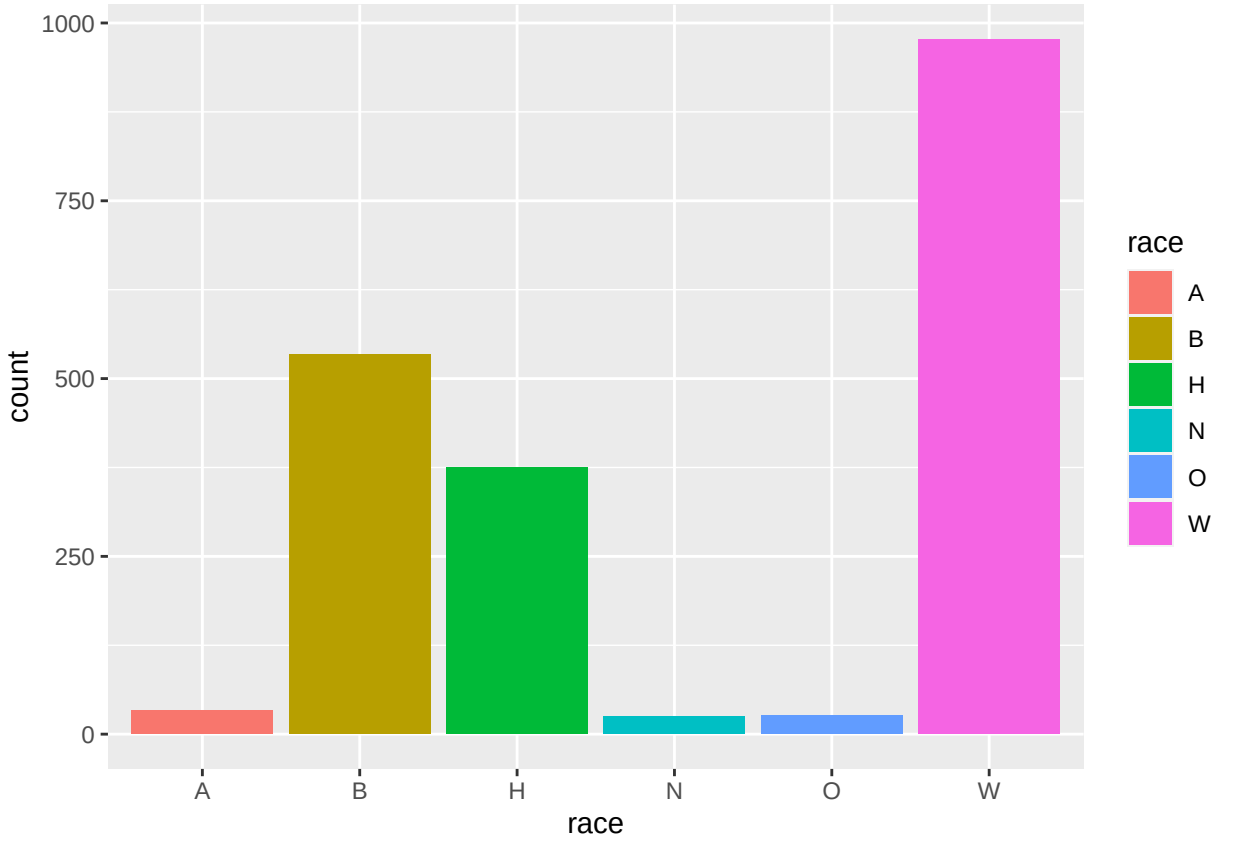


Podem destacar que, clarament, la gran majoria de persones disparades són homes.

Distribució d'ètnies

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

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



De qui més registres tenim és de les persones blanques. És lògic veure aquests resultats, ja que la majoria de persones en Estats Units són de blanques, però, quina és la proporció de víctimes de cada respectiva ètnia respecte al total de persones d'eixa ètnia?

així que s'ajustaran bastant bé.

```
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"
```

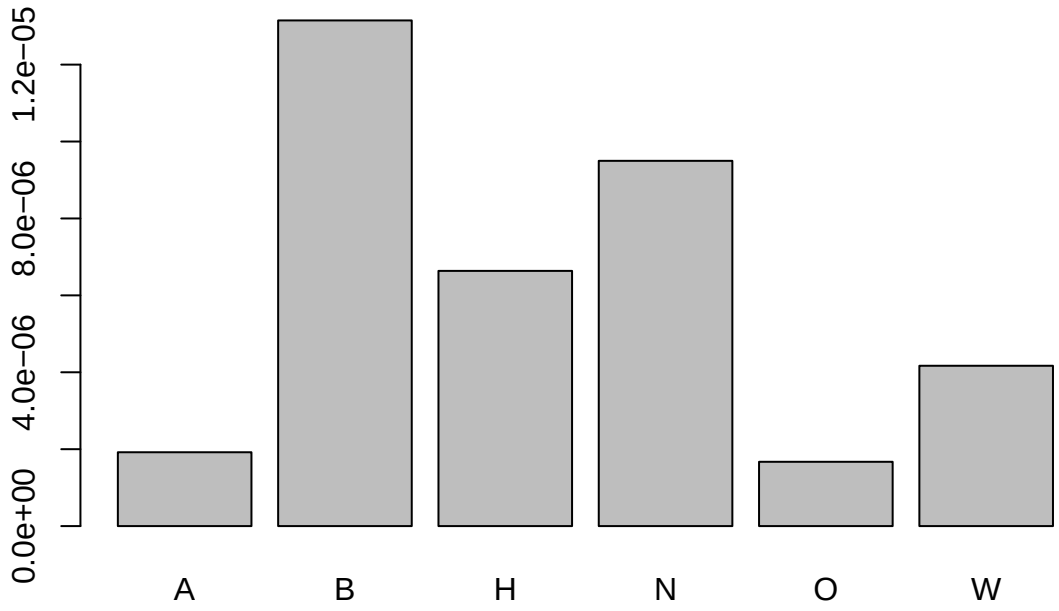
La data mínima de les nostres dades és 02/01/2015 i la màxima, 31/07/2017.

Vegem el gràfic de les víctimes de tirs tenint en compte la població total de cada ètnia:

```
races_total <- c((sum(df_clean$race == 'A') / 17186320), (sum(df_clean$race == 'B') / 40610815), (sum(d

# Etiquetes
races <- c('A', 'B', 'H', 'N', 'O', 'W')

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



Veient el gràfic amb les dades comparades amb el total de població de cada ètnia pareix indicar que les persones negres, natives americanes i hispanes són més probables de rebre un tir d'un policia que una persona blanca.

Poverty i high school per ètnies

Per analitzar un poc mes en profunditat algunes possibles relacions entre les persones que han sigut disparades, discretitzarem les variables percent_completed_hs i poverty_rate. Dividirem els seus valors en les franges: 0-24, 25-49, 50-74 i 75-100.

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

Vegem la distribució de les dades en aquestes franges:

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

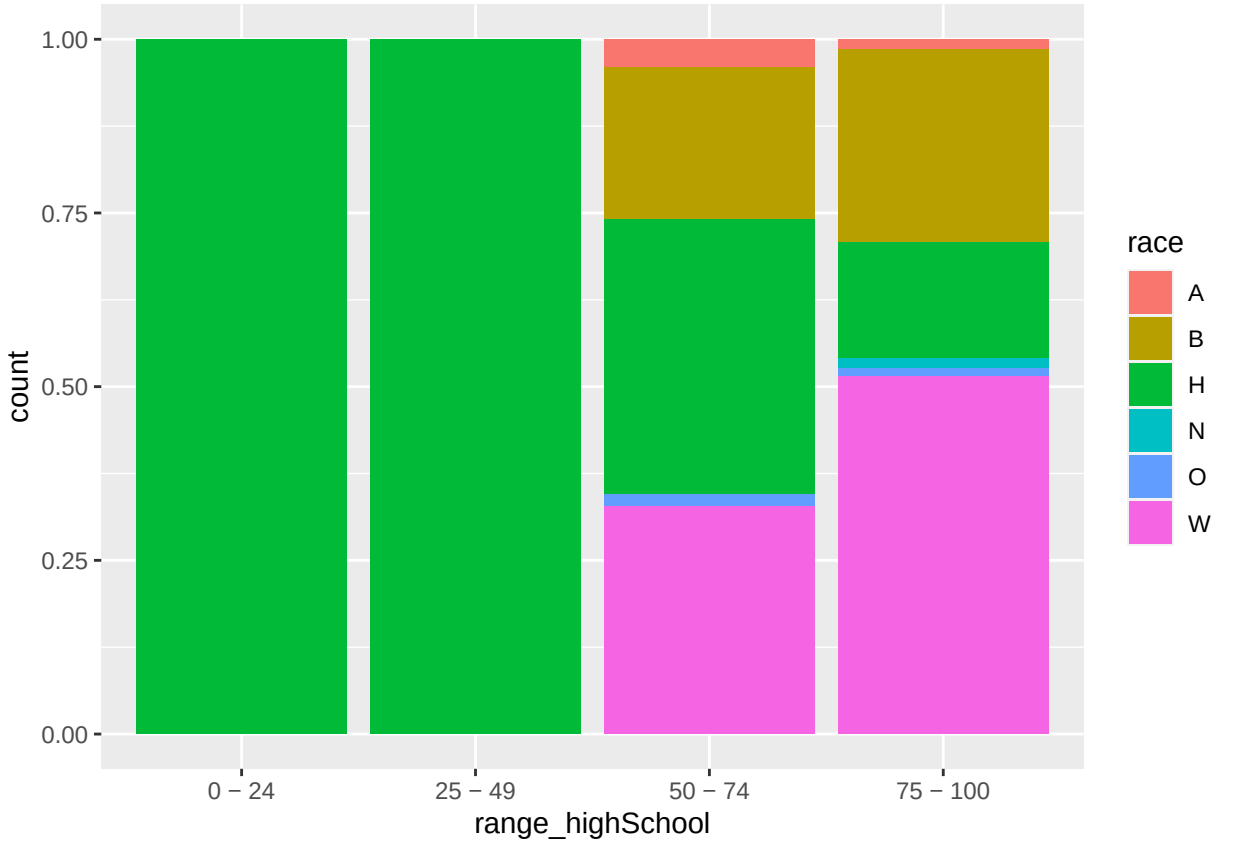
```
##
##   0 - 24  25 - 49  50 - 74 75 - 100
##        1        6      174     1789
```

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

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

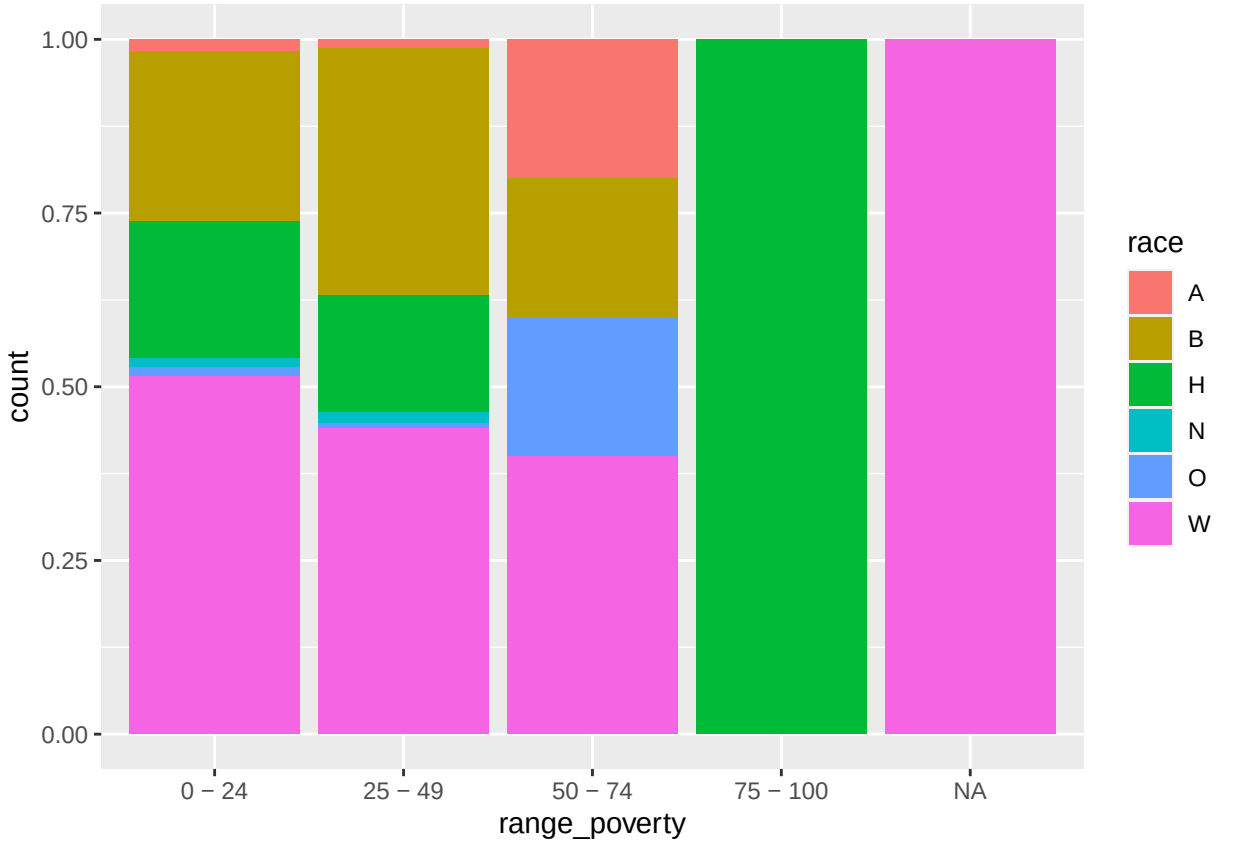
Analitzem-ho gràficament:

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



Ens fixem en 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")
```



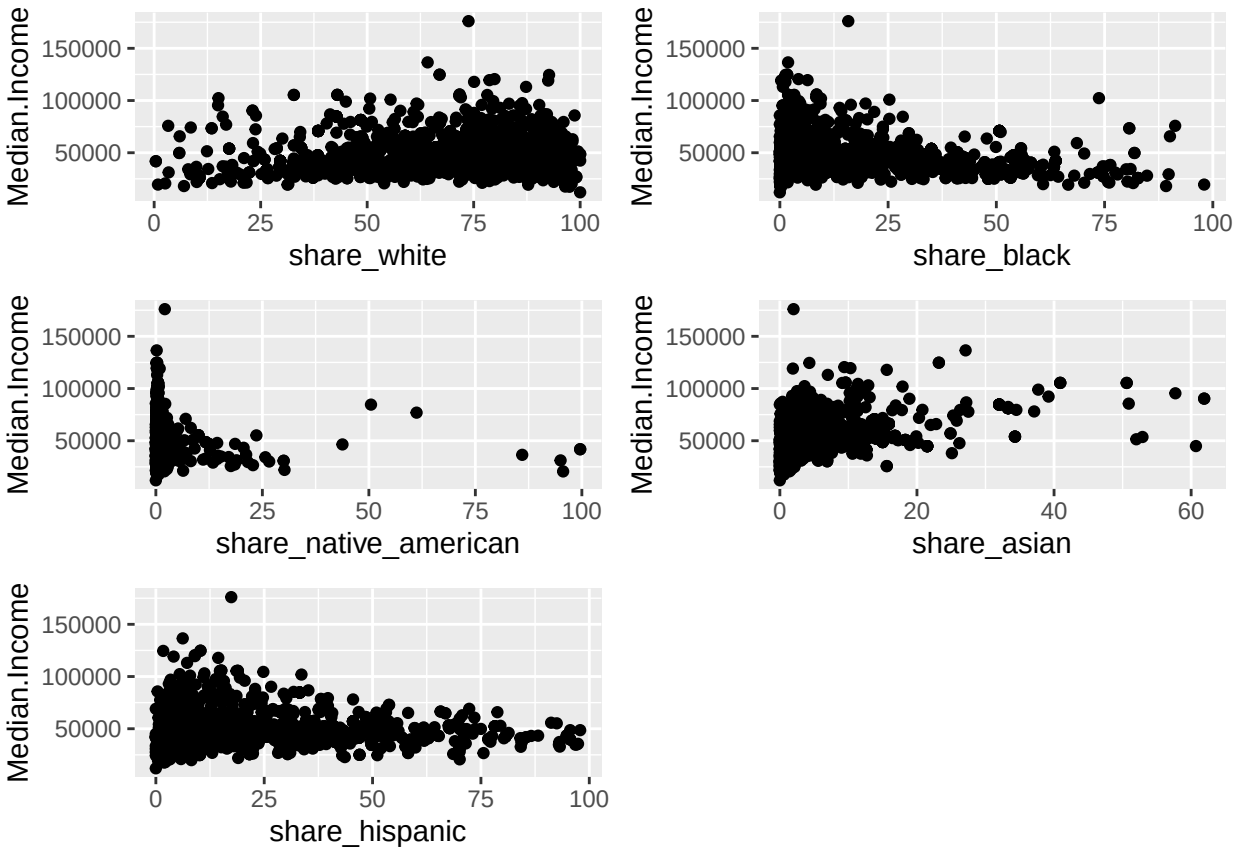
En aquest gràfic ens ficarem en les franges de 0 a 50%. Les altres franges no tenen suficients dades per a ser significants. Pareix que hi ha una diferència notable entre les persones blanques i negres. El percentatge de les persones blanques amb rang de pobresa entre el 0 i 25% és superior al de les persones blanques en el rang 25-50%. Amb les persones negres ocorre el contrari, hi ha un percentatge major de persones en rang de pobresa entre el 25-50% que entre el 0 i 25%.

Median Income per ètnies

En aquest apartat compararem la distribució de l'ingrés mitja de les ciutats tenint en compte la representació de cada ètnia.

```
plot_share_white <- ggplot(df_clean) + geom_point(map = aes(x = share_white, y = Median.Income))
plot_share_black <- ggplot(df_clean) + geom_point(map = aes(x = share_black, y = Median.Income))
plot_share_native_american <- ggplot(df_clean) + geom_point(map = aes(x = share_native_american, y = Me
plot_share_asian <- ggplot(df_clean) + geom_point(map = aes(x = share_asian, y = Median.Income))
plot_share_hispanic <-ggplot(df_clean) + geom_point(map = aes(x = share_hispanic, y = Median.Income))

grid.arrange(plot_share_white, plot_share_black, plot_share_native_american, plot_share_asian, plot_sha
```



Com és lògic les ciutats amb un alt percentatge de persones blanques són les que més tenim.

Pel que fa a la comparació en els ingressos en ciutats amb un percentatge molt elevat de cada ètnia podem veure certes diferències. Les ciutats amb alt percentatge de persones blanques són les que més alts ingressos presenten, encara que també hi ha ciutats amb alt percentatge de persones negres o asiàtiques amb mitjans-alts ingressos. En canvi, en ciutats amb alt percentatge de persones hispanes o natives americanes no hi tenen uns grans ingressos.

En els següents apartats ens centrarem en les persones blanques, negres i hispanes per a analitzar-les més, ja que són les que més dades tenim.

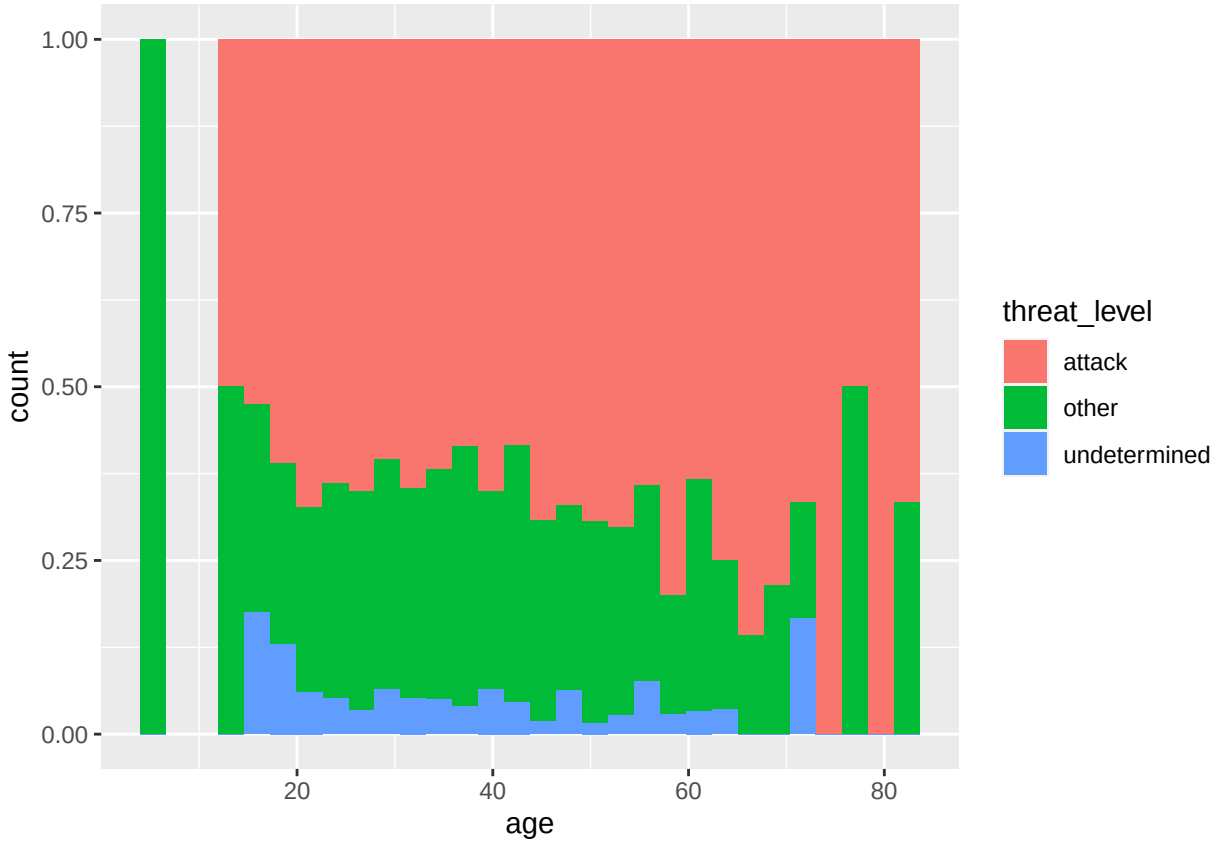
Amenaça i manera en què ha mort la persona

Per últim, mitjançant histogrames vegem com es distribueixen els percentatges dels tipus d'amenaça que s'indiquen i la manera en què ha mort la persona relacionant-ho amb l'edat:

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

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

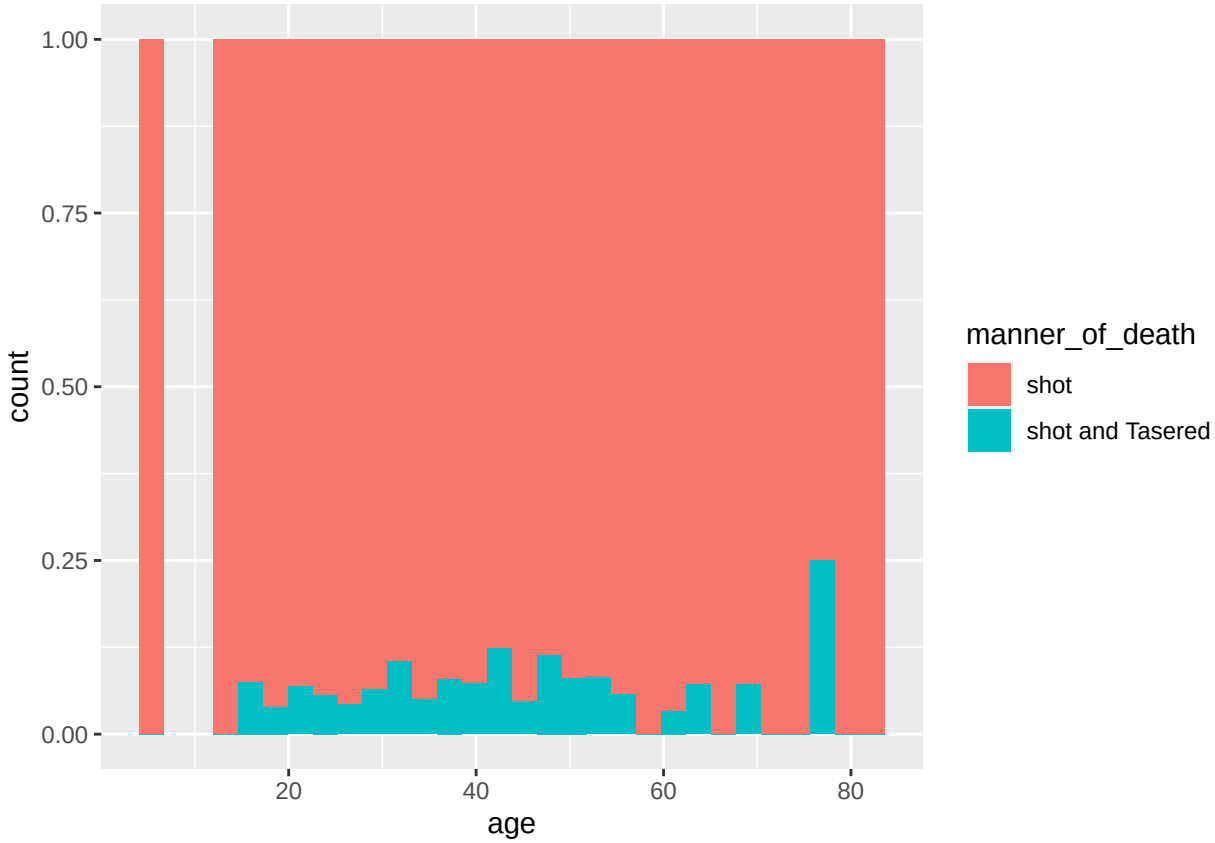
```
## Warning: Removed 6 rows containing missing values (geom_bar).
```




```
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).
```



No pareix haver-hi una diferència notable entre el tipus d'amenaça i l'edat. La majoria es consideren atacs en la mateixa proporció.

I tampoc pareix que canvia la proporció de persones mortes per un tir o un tir i Taser amb l'edat. Tenim un pic al voltant dels 80 anys, però és possible que siga degut al fet que en eixes edats tinguem molt poques dades.

Regressió

A partir de les dades ja tractades obtenim un dataset amb la informació respectiva a les ciutats.

Contindrà les variables:

- state
- city
- percent_completed_hs
- poverty_rate
- Median.Income
- share_white
- share_black
- share_native_american
- share_asian
- share_hispanic

```
df_clean_cities <- unique(df_clean[,c(1,2,14,15,16,17,18,19,20,21)])
```

A partir d'aquestes dades construirem tres models de regressió múltiple que ens podrien ajudar a predir informació sobre alguna ciutat a partir d'altres dades:

Model 1: Predicció de l'ingrès mitja a partir de les variables share_white, share_black.

```
model_1 <- lm(formula = Median.Income ~ share_white + share_black + share_hispanic + percent_completed_
summary(model_1)
```

```
##
## Call:
## lm(formula = Median.Income ~ share_white + share_black + share_hispanic +
##     percent_completed_hs + poverty_rate, data = df_clean_cities)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47690  -5789  -1475   3380 103267
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          52990.47    6018.62   8.804  < 2e-16 ***
## share_white           -282.52      27.16 -10.404  < 2e-16 ***
## share_black           -201.25      31.83  -6.324 3.67e-10 ***
## share_hispanic         119.68      24.78   4.830 1.55e-06 ***
## percent_completed_hs   483.43      54.48   8.874  < 2e-16 ***
## poverty_rate         -1281.58      46.58 -27.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10700 on 1128 degrees of freedom
## Multiple R-squared:  0.6381, Adjusted R-squared:  0.6365
## F-statistic: 397.8 on 5 and 1128 DF,  p-value: < 2.2e-16
```

Model 2: Predicció de la variable poverty_rate amb share_white, share_black i share_hispanic.

```
model_2 <- lm(formula = poverty_rate ~ share_white + share_black + share_hispanic, data = df_clean_citi
summary(model_2)
```

```
##
## Call:
## lm(formula = poverty_rate ~ share_white + share_black + share_hispanic,
##     data = df_clean_cities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.482  -5.581  -0.973   4.866  55.513
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.650552   1.973459   7.424 2.23e-13 ***
## share_white     0.002626   0.021304   0.123    0.902
## share_black     0.191992   0.024256   7.915 5.86e-15 ***
## share_hispanic  0.085849   0.015579   5.511 4.42e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.399 on 1130 degrees of freedom
## Multiple R-squared:  0.1481, Adjusted R-squared:  0.1458
## F-statistic: 65.46 on 3 and 1130 DF,  p-value: < 2.2e-16
```

Model 3: Predicció de el percentatge de persones que han superat els estudis superriors a partir de les variables share_white, share_black i share_hispanic.

```
model_3 <- lm(formula = percent_completed_hs ~ share_white + share_black  + share_hispanic, data = df_cl
summary(model_3)
```

```
##
## Call:
## lm(formula = percent_completed_hs ~ share_white + share_black +
##     share_hispanic, data = df_clean_cities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.511  -3.611   1.210   4.817  19.616
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    92.57911    1.68722  54.871  < 2e-16 ***
## share_white    -0.01698    0.01821  -0.932    0.351
## share_black    -0.13597    0.02074  -6.557 8.35e-11 ***
## share_hispanic -0.30826    0.01332 -23.144  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.181 on 1130 degrees of freedom
## Multiple R-squared:  0.4165, Adjusted R-squared:  0.415
## F-statistic: 268.9 on 3 and 1130 DF,  p-value: < 2.2e-16
```

Clustering

En aquest punt anem a agrupar el dataset de df_clean_cities en clusters per veure si les ciutats es podrien agrupar segons algunes característiques.

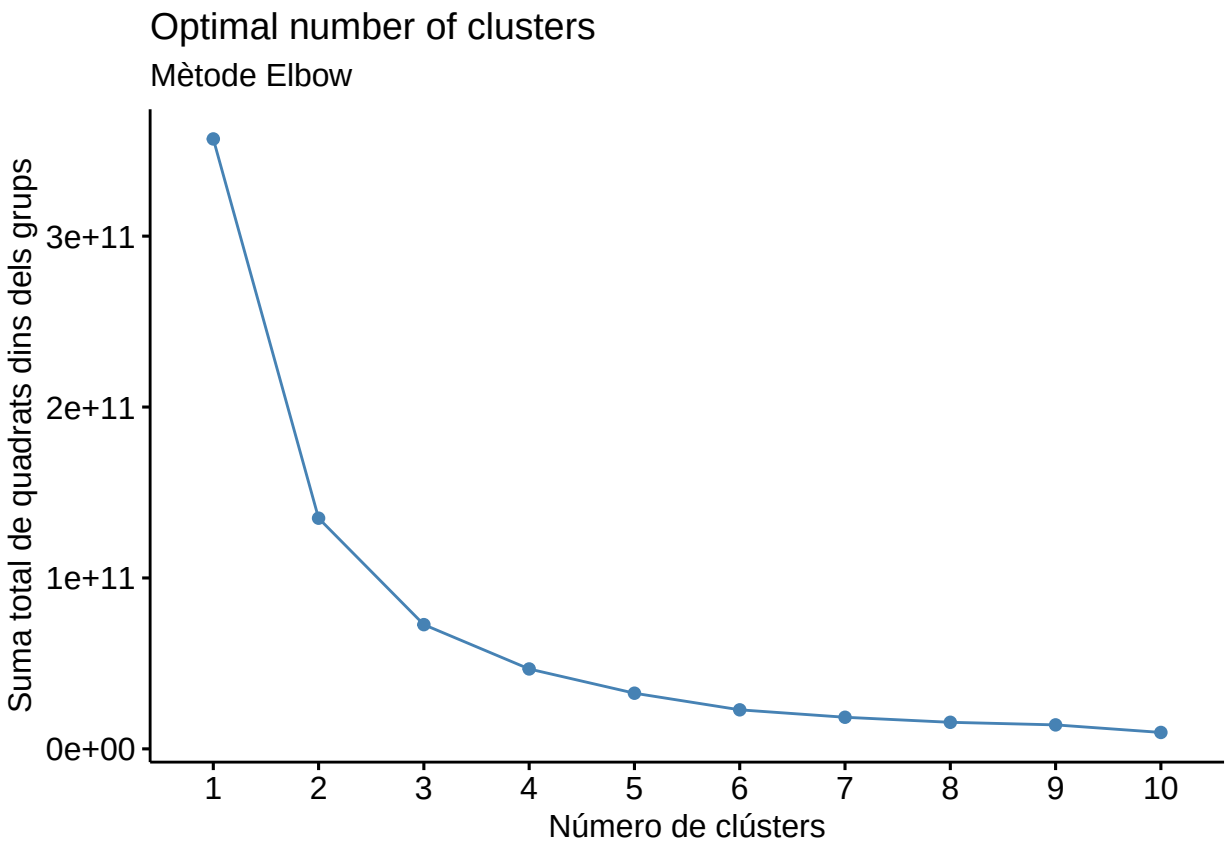
Aplicarem l'algoritme kmeans. Per saber quin k serà el mes òptim farem servir el mètode del colze amb l'ajuda de la següent funció:

```
library(cluster)
library(fpc)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.5
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_nbclust(df_clean_cities[3:10], kmeans, method = "wss") +
   labs(subtitle = "Mètode Elbow") +
   xlab("Número de clústers") +
   ylab("Suma total de quadrats dins dels grups")
```



Contrast d'hipòtesis

En l'analisi preliminar de les dades hem vist que la distribució d'edats per ètnia no es igual per a totes. Les persones blanques assasinades solen ser mes majors que la resta. Però aquesta diferència es significant?

Per a saber-ho farem un contrast d'hipotesi entre les edats de les persones blanques i les persones negres.

Les hipòtesis seràn les següents: - Nula mitjanes edat igual - Alternativa mitjanes edat distint

Comprovacio suposicions

Realitzacio contrast

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