**Pràctica 2: Neteja i validació de les dades**

Eva Suñé Expósito

11 de Juny del 2018

# **Índex**

**1. Descripció i objectius del dataset** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

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

**3. Neteja de les dades**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

3.1. Identificació i tractament d’elements buits o zeros. . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

3.2. Identificació i tractament de valors extrems. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

**4. Anàlisi de les dades**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

4.1. Selecció dels grups de dades que es volen analitzar/comparar. . . . . . . . . . . . . . . . . .5

4.2. Comprovació de la normalitat i homogeneïtat de la variància. . . . . . . . . . . . . . . . . . .7

4.3. Aplicació de proves estadístiques per comparar els grups de dades. . . . . . . . . . . . . 8

4.3.1. Correlació entre les diferents variables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

4.3.2. Proves de contrast d’hipòtesis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

4.3.3. Model de regressió lineal. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

4.3.4. Model de regressió logística. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10

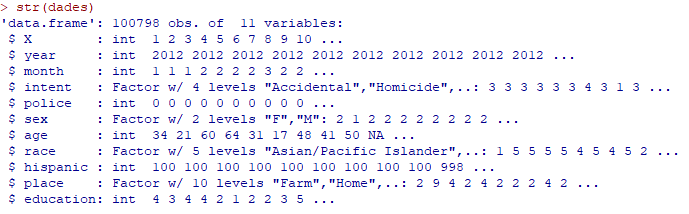
**6. Resolució del problema**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

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

El conjunt de dades objecte d’anàlisis s’ha obtingut de la pàgina web Kaggle (<https://www.kaggle.com>), el dataset seleccionat s’anomena: **Gun Deaths in the US: 2012-2014**. Aquest fitxer consta de totes les morts per arma de foc als Estats Units entre el 2012 i el 2014.

El dataset esta format per 11 variables (columnes) que presenten 100798 víctimes (ﬁles o registres). Els diferents camps d’aquest conjunt de dades són:

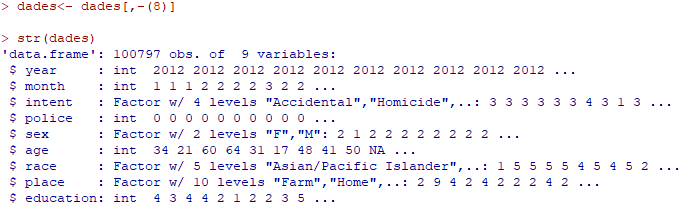
* **X:** número d’identificació del registre.
* **Year:** any del succés.
* **Mounth:** mes del succés.
* **Intent:** la intenció de l’autor del crim, es cataloga en: Suicide, Accidental, NA, Homicide, o Undetermined
* **Police:** si un agent de policia es va involucrar en el tiroteig, 0 (Fals) o 1 (Cert).
* **Sex:** gènere de la víctima, M (Home) o F (Dona).
* **Age:** edat de la víctima.
* **Race:** l’ètnia de la víctima, Asian/Pacific Islander, Native American/Native Alaskan, Black, Hispanic, o White.
* **Hispànic:** Un codi que indica l'origen hispànic de la víctima.
* **Place:** lloc del succés
* **Education:** situació educativa de la víctima. Es classifica en diferents opcions:
  + 1: menys que l'escola secundària
  + 2: es va graduar a l'escola secundària o equivalent
  + 3: Universitari
  + 4: com a mínim es va graduar a la Universitat
  + 5: no disponible

En aquest conjunt de dades es pot obtindré informació sobre els diferents perfils de víctimes, segons edat, sexe, ètnia....

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

Els atributs del conjunt de dades es corresponen a les característiques de les víctimes, per tant serà convenient tenir-los en consideració. Hi ha dos atributs que no aporten informació sobre les víctimes, un és la primera columna, que són el ID dels registres i el segon és l’atribut hispànic, per aquesta raó els eliminarem.





**3. Neteja de les dades.**

**3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?**

En aquest dataset els zeros no implica falta de valor, ja que si que té significat en la majoria dels atributs, així que per indicar que no hi ha valor en el registre s’utilitza NA. A continuació, mostrem quins camps contenen elements buits:



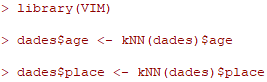
Arribats aquest punt, s’ha de decidir com actuar sobre aquest registres que no contenent informació, com hi ha diferents atributs que tenen valors buits es realitzarà diferents accions, depenen de cada atribut i el número de valors buits.

Una opció podria ser eliminar els registres que continguin el valor NA, però això suposarà perdre informació. En el cas de l’atribut intent, si que eliminarem l’únic registre sense valor, ja que la pèrdua d’informació és casi mínima.





Un altre alternativa, és utilitzar el mètode d’imputació de valors basat en la similitud o diferencia entre els registres: la imputació es basa en k veïns més propers (en angles, kNN-imputation). L’elecció d’aquesta alternativa es realitza sota la hipòtesis de que els nostres registres guarden certa relació. No obstant, es millor treballar amb dades “aproximades” que amb elements buits, ja que obtindrem un anàlisis amb menys marge d’error.



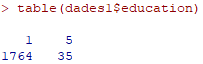


En el cas de l’atribut education abans de realitzar el mètode kNN-imputation filtrarem per totes les víctimes que siguin menors de 16 anys, en aquest cas l’atribut education hauria de ser 1:





A més, si es revisen els valors de education s’observen valors erronis:



Per aquest motiu reemplacem tots els 5 i NA pel valor 1.





Ara aplicarem el mètode kNN-imputation.

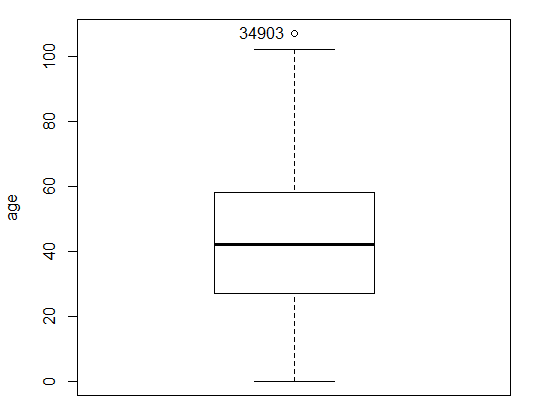




**3.2. Identificació i tractament de valors extrems.**

Els valors extrems o outliers són aquells que semblen no tenir sentit si els comparem amb la resta de dades. Per identiﬁcar-los, es pot fer ús de dos vies: (1) utilitzar la funció boxplots.stats() de R o (2) representar un diagrama de caixa per cada variable i veure quins valors queden fora del rang interquartílic (la caixa). En el nostre conjunt de dades l’únic atribut que té sentit calcular els valors extrems és el camp age, per aquest motiu representarem els outliers pels dos mètodes.





Observem que només hi ha un valor (107) que estigui fora de rang, el resultat és viable ja que pot ser, així que no es realitzarà cap acció amb els valors extrems.

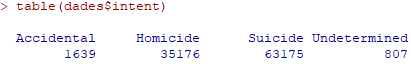
**4. Anàlisi de les dades.**

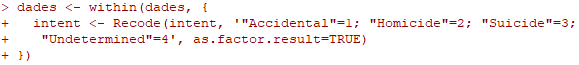
**4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).**

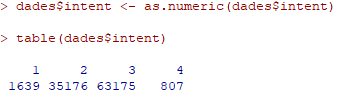
Per poder analitzar les dades correctament assignem a les variables qualitatives un valor.

**Atribut intent:**

* Accidental → 1
* Homicide → 2
* Suicide → 3
* Undetermined → 4

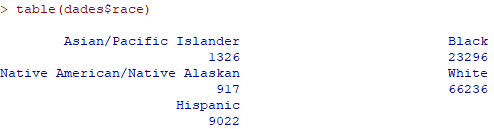


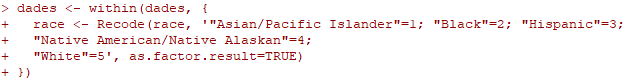




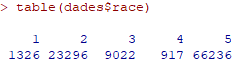
**Atribut race:**

* Asian/Pacific Islander → 1
* Black → 2
* Hispanic → 3
* Native American/Native Alaskan → 4
* White → 5



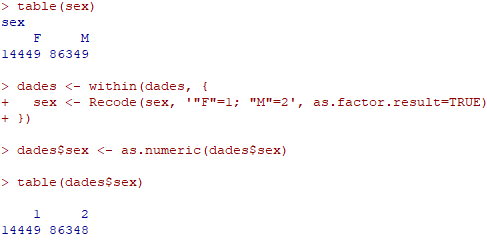




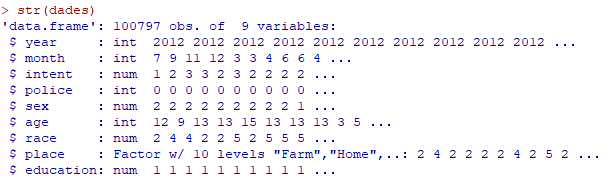


**Atribut sex:**

* F → 1
* M → 2



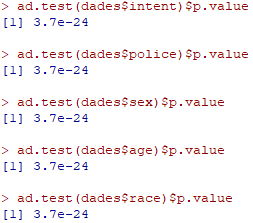
L’únic atribut qualitatiu que no canviem a un valor quantitatiu és l’atribut place, ja que hi ha molts valors. El resultat final és:



**4.2. Comprovació de la normalitat i homogeneïtat de la variància.**

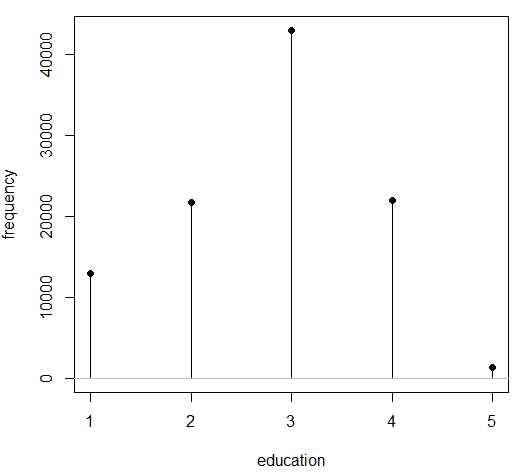
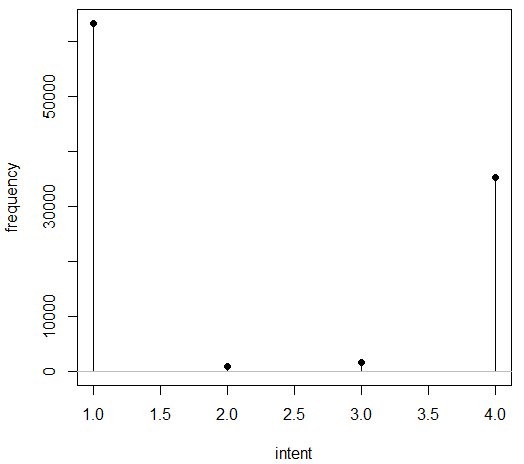
Totes les variables són variables discretes per veriﬁcar si compleixen l’assumpció de normalitat utilitzarem la prova d’AndersonDarling.



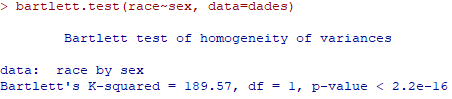




No hi ha cap variable que segueixi una distribució de normal. A continuació es mostra dues de les variables triades a l’atzar per mostrar la seva distribució:

Per revisar la homogeneïtat de la variància utilitzarem el test de bartlett, ja que les mostres no són normals. En aquest cas, estudiarem la homogeneïtat entre les diferents ètnies i el sexe de les víctimes. En el següent test, la hipòtesis nul·la consisteix en que les dues variàncies són iguals.

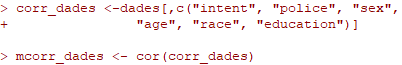


Com el p-valor és inferior 0,05, aleshores les variàncies no són homogènies.

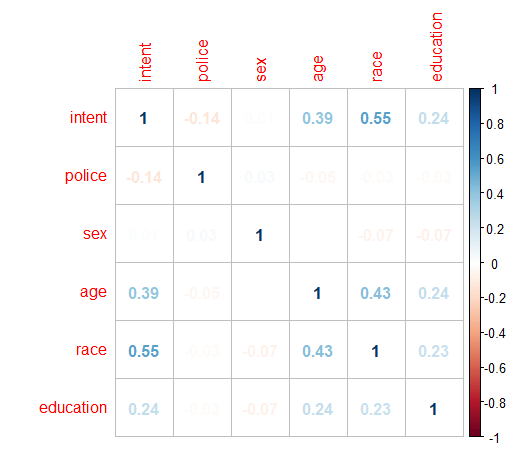
**4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l’objectiu de l’estudi, aplicar proves de contrast d’hipòtesis, correlacions, regressions, etc.**

**4.3.1. Correlacions entre les diferents variables**

Calcularem si hi ha alguna relació entre les diferents variables, per això calculem la matriu de correlació.







Es pot observar com hi ha una relació entre la ètnia de la víctima amb l’edat i també hi ha relació amb com es va produir la mort de la víctima (suïcidi, assassinat, accident o indeterminat). També veiem que la policia o el sexe de la víctima no tenen incidència en el succés.

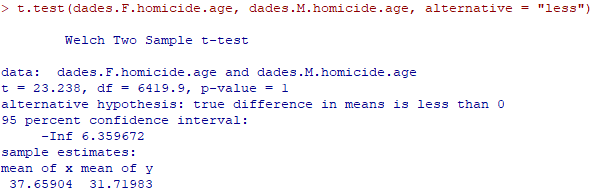
**4.3.2. Proves de contrast d’hipòtesis**

La segona prova estadística que aplicarem consistirà en un contrast de hipòtesis sobre dos mostres per determinar si en els homicidis on les dones són més grans que els homes. Per realitzar la prova utilitzarem dos mostres diferents: la primera correspon a les víctimes d’homicidi que són dones i la seva edat i, la segona, les víctimes d’homicidi masculines i l’edat.



Es planteja el següent contrast d’hipòtesis de dos mostres sobre la diferencia de les mitjanes, el qual és unilateral atenen a la formulació de la hipòtesis alternativa:

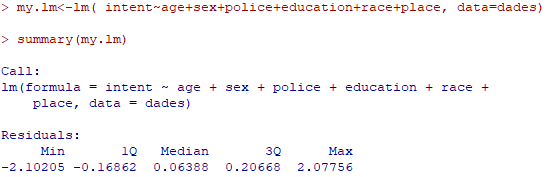
H0 : µ1−µ2 =0 H1 : µ1−µ2 < 0

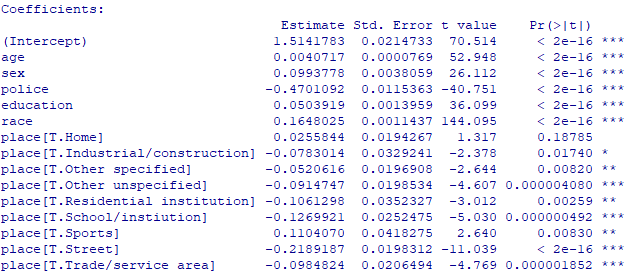


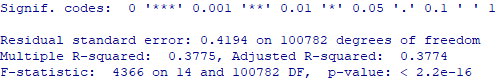
Com que obtenim un p-valor major al valor de signiﬁcació ﬁxat, α =0,05, acceptarem la hipòtesis nul·la. Per tant, podrem afirmar que l’edat de les víctimes d’homicidi de les dones és superior a l’edat dels homes.

**4.3.3. Model de regressió lineal**

Un dels objectius de l’activitat és poder predir quin col·lectiu és més propens de tenir un assassinat, suïcidi o accident donades les característiques de les persones. Per tant, es calcularà un model de regressió lineal utilitzant regressors tant quantitatius como qualitatius amb els que poder realitzar les prediccions de quin perfil de persona és més propensa a un assassinat, suïcidi o accident.







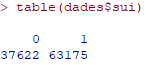
Amb el model de regressió lineal s’obté un coeﬁcient de la bondat de l’ajust del 0.3775, aquest valor és dolent per poder obtenir una predicció eficient.

**4.3.4. Model de regressió logística**

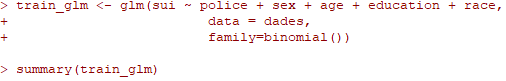
Per poder realitzar un model de regressió logística serà necessari que la variable dependent sigui una variable binària. En aquest cas ens centrarem amb les víctimes que es van suïcidar. Així que les víctimes que es van suïcidar tindran el valor 1 i la resta 0, crearem una variable nova: sui.

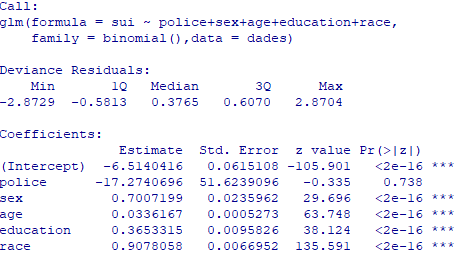


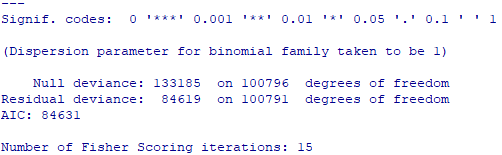




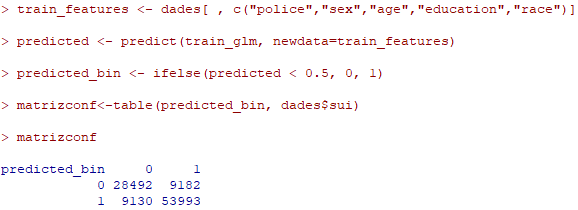
Un cop tenim les dades en el format correcte, creem el model.







Ara avaluarem el model amb les mateixes dades del nostre dataset.

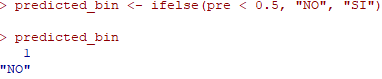




Amb aquest model obtenim una qualitat de l’ajust del 0.818.

Creem un nou dataset amb els següents paràmetres per avaluar si té risc de suïcidar-se.





Amb aquest paràmetres, el resultat obtingut és que no té possibilitats que es suïcidi.

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

Com s’ha observat, hem realitzat quatre probes d’estadístiques diferents sobre el conjunt de dades que es corresponia amb les diferents víctimes per arma de foc a EEUU amb les seves característiques. Per cada una de elles, hem visualitzat els diferents resultat que ens plantejàvem i els coneixements que hem pogut extreure mitjançant els diferents anàlisis. L’anàlisi de correlació i el contrast de hipòtesis ens ha permès conèixer quines variables tenien més importància sobre la intencionalitat de la víctima, mentre que els models de regressió lineal i regressió logística s’han pogut realitzar prediccions a partir de les variables, encara que el primer model no hem obtingut una bona qualitat alhora de predir la intencionalitat sobre diferents persones, amb el segon model si que hem obtingut una bona qualitat per predir.

Abans de poder analitzar les dades, s’ha hagut de realitzar un pre-processament i neteja d’elles. Un dels pre-processament que hem realitzat és eliminar els valors buits, en el nostre cas hem realitzat diferents mètodes. També hem avaluat els valors extrems, en aquest cas hem optat per no realitzar cap acció, ja que l’únic valor que hem apreciat no ens afectava alhora d’analitzar les dades.