Pràctica 2

1. Descripció del dataset.

1. El dataset.

El conjunt de dades s'ha obtingut a través del portal Kaggle. El dataset presenta els accidents registrats pels agents de Policia durant el 2016 a la ciutat de Barcelona. Concretament, defineix un total de 12.072 accidents mitjançant 28 columnes que detallen les diferents característiques dels successos.

A continuació es detallen els camps del dataset:

- *Número d'expedient:* Número d'identificació de l'expedient
- *Codi districte:* Codi del districte.
- *Nom districte:* Nom del districte.
- *Codi barri:* Cadi del barri.
- *Nom barri:* Nom del barri.
- *Codi carrer:* Codi del carrer.
- *Nom carrer:* Nom del carrer.
- *Num postal caption:* Número postal.
- *Descripció dia setmana:* Dia de la setmana.
- *Dia setmana:* Diminutiu del dia de la setmana.
- *Descripció tipus dia*: Tipus de dia ("laboral" o "festiu").
- *NK Any:* Any.
- *Mes de any:* Mes de l'any.
- *Nom mes:* Nom del mes.
- *Dia de mes:* Dia del mes.
- *Descripció torn:* Torn del policia. ("Matí", "Tarda", "Nit")
- *Hora de dia*: Hora del dia (Format: 24h)
- *Descripció causa vianant:* Descripció de la causa, si s'escau.
- *Desc. Tipus vehicle implicat:* Descripció del tipus de vehicle implicat
- *Descripció sexe:* Sexe de l'implicat.
- *Edat*: Edat de la víctima.
- *Descripció tipus persona*: Descriu el tipus de víctima ("Conductor", "Passatger", "Vianant").
- *Descripció situació:* Descripció de la situació ("Desconegut", "Otros", "Presentado")
- *Descripció victimització:* Descripció del resultat de les víctimes ("Ferit lleu", "Ferit greu", "Mort")

- *Coordenada UTM (Y):* Coordenada Y en format UTM
- Coordenada UTM (X): Coordenada X en format UTM
- *Long:* Longitud
- *Lat:* Latitud

2. Importància i objectius de l'anàlisi.

La seguretat viària és un tema important per la ciutat de Barcelona, de manera que l'estudi del registre d'aquests fets pot aportar-nos informació que no es veu a simple vista. És a dir, a partir d'aquests dataset es planteja definir quins són els atributs que més influeixen a l'hora de patir un accident de trànsit a la ciutat.

Totes les grans ciutats es marquen com a objectiu reduir la quantia d'accidents. D'aquí la importància d'estudiar els accidents passats per poder, de certa manera, prevenir-los en un futur. De manera paral·lela, l'estudi d'aquest registre pot ser clau en la definició de situacions en les quals la seguretat viaria es veu compromesa.

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

A causa de la riquesa del dataset, no ha sigut necessari realitzar cap integració en el mateix. D'altra banda, si que s'ha dut a terme una selecció d'atributs d'interès. Ja que molts d'ells expressaven o bé valors no útils per l'estudi o bé informació repetida en un format distint, com per exemple el nom dels mesos i el seu respectiu valor numèric. Per aquesta raó s'ha decidit eliminar els següents atributs, especificant el motiu entre parèntesis:

- *Número d'expedient* (Dada no representativa en l'estudi)
- *Codi districte* (Dada no representativa en l'estudi)
- *Nom districte* (Dada no representativa en l'estudi)
- *Codi barri* (Dada no representativa en l'estudi)
- *Nom barri* (Dada no representativa en l'estudi)
- *Codi carrer* (Dada no representativa en l'estudi)
- *Nom carrer* (Dada no representativa en l'estudi)
- *Num postal caption* (Dada no representativa en l'estudi)
- *Descripció dia setmana* (dada redundant)
- *Descripció tipus dia* (Totes les dades s'han recopilat en dies laborals)
- *NK Any* (El dataset fa referència únicament a l'any 2016)
- *Nom mes* (dada redundant)
- *Descripció torn* (dada redundant)

- *Descripció causa vianant*
- *Descripció situació* (Dada no representativa en l'estudi)
- *Coordenada UTM (Y)* (Dada no representativa en l'estudi)
- Coordenada UTM (X) (Dada no representativa en l'estudi)
- *Long* (Dada no representativa en l'estudi)
- *Lat* (Dada no representativa en l'estudi)

3. Neteja de les dades.

1. Gestió de dades buides.

Per fer la recerca de les dades que contenen zeros o elements buits s'ha fet ús de la funció "`table`". Aquesta funció permet obtenir un recompte dels valors presents en un atribut. S'han detectat únicament dos atributs (`Descripció sexe`, `Edat`) que contenen dades errònies. A continuació es mostra el resultat de l'execució de la funció per ambdós atributs:

```
> table(Accidents$`Edat`)

      0       1      10      11     119      12      13
     24      19      33      31       1      40      27
     14      15      16      17      18      19       2
     36      37      48      67     124     204      25
     20      21      22      23      24      25      26
    214     218     294     358     338     339     363
     27      28      29       3      30      31      32
    338     346     319      29     364     318     278
     33      34      35      36      37      38      39
    308     297     288     289     281     291     262
      4      40      41      42      43      44      45
     22     251     286     234     230     246     230
     46      47      48      49       5      50      51
    210     206     202     193      31     190     198
     52      53      54      55      56      57      58
    179     158     149     133     112     116     108
     59       6      60      61      62      63      64
     94      20      91      88      73      69      61
     65      66      67      68      69       7      70
     47      62      47      42      40      28      50
     71      72      73      74      75      76      77
     35      37      38      40      32      27      27
     78      79       8      80      81      82      83
     28      37      24      28      33      28      39
     84      85      86      87      88      89       9
     21      24      21      16      16       8      27
     90      91      92      93      94      96 Desconegut
      8       5       9       3       5       1     111
> table(Accidents$`Descripció sexe`)

Desconegut       Dona       Home
         1       4746       7325
```

Figura 1. Ús de la funció `'table'` als atributs Edat i Descripció sexe.

Pel que fa a la taula "`Edat`", es consideraran com a dades errònies ambdós valors i a continuació es detallen els motius. Considerant la magnitud del dataset (12072 observacions), es pot afirmar que l'eliminació de 135 observacions només representa un 1,12% de la mostra, i per tant la seva absència no és representativa. Deixant de banda els valors percentuals, s'ha de

destacar que tenir 0 anys només es pot donar en el cas que siguin dues víctimes i una d'elles estigui embarassada. Però al no disposar del recompte de víctimes de cada accident, no es pot avaluar aquesta hipòtesis i per tant es descarta. A banda d'això, en aquest cas les dades no poden ser interpolades a partir de les dades disponibles, ja que l'edat no és una dada que puguem deduir a partir del dataset.

De manera paral·lela, la taula "*Descripció sexe*" conté un sol valor desconegut que, seguint el raonament anterior, es procedirà a l'eliminació de l'observació.

```
AccidentsOK <- AccidentsOK[-c(which(AccidentsOK$Edat == '0')),]
AccidentsOK <- AccidentsOK[-c(which(AccidentsOK$Edat == 'Desconegut')),]
```

Figura 2. Funció utilitzada per la gestió de les dades errònies.

Finalment, es procedirà a la normalització de les variables qualitatives en format text per obtenir variables numèriques però sense valor numèric, en el estricte sentit de la paraula. Pel que fa a la normalització, es durà a terme mitjançant les funcions "replace" i "which", que conjuntament permeten localitzar i substituir els valors originals pels valors d'interès.

```
AccidentsOK$`Descripció sexe` <- replace(AccidentsOK$`Descripció sexe`, which(AccidentsOK$`Descripció sexe` == "Dona"), 1 )
```

Figura 3. Exemple de codi utilitzat per la normalització de les dades.

Aquesta operació es durà a terme en els atributs següents:

- *Dia setmana*
- *Desc. Tipus vehicle implicat*
- *Descripció sexe*
- *Descripció tipus persona*
- *Descripció victimització*

La taula que és mostra en l'annex 1 servirà de referència per a la conversió.

2. Gestió de valors extrems.

Pel que fa als "outliers" o valors extrems, la seva identificació s'ha dut a terme mitjançant un diagrama de caixa *(funció `boxplot()`)* per a cada columna del dataset, la figura 4 en mostra un exemple. Convé destacar que només s'han detectat "outliers" a l'atribut 'Edat' i que cap altra columna del dataset presenta valors extrems. La figura 4 permet apreciar que l'atribut 'Edat' conté diferents punts atípics en el límit superior del diagrama.

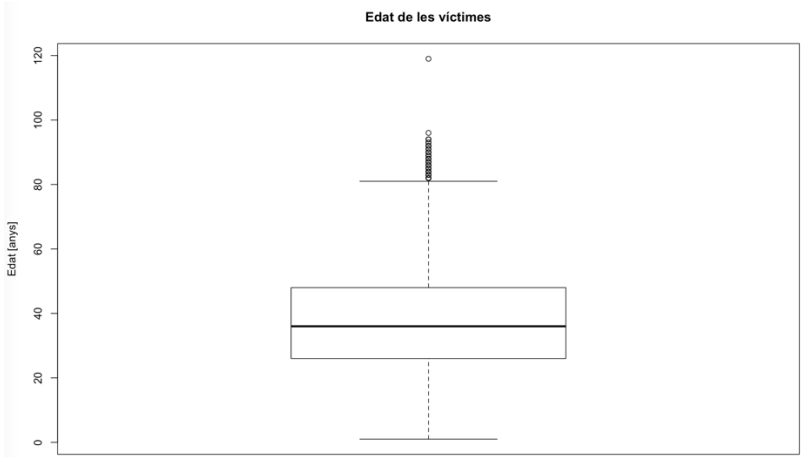

Figura 4. Diagrama de caixa de l'atribut 'Edat'

Tenint en consideració que una persona sana pot viure entre 75 i 100 anys, només s'observa un "outlier" que frega els 120 anys. Totes les altres observacions es poden donar com a vàlides donat que entren dins el ventall d'edat real menys el punt que fa referència a una víctima de 119 anys. Conseqüentment, es passa a considerar l'observació com un punt erroni i es procedeix a la seva eliminació, per les següents raons:

- El rècord mundial de longevitat està establert a 114 anys.
- És molt poc probable que una persona d'edat tan avançada estigui implicada un accident de circulació.

A tall de conclusió, després de realitzar les tasques de neteja de les dades disposem d'un dataset format per 9 atributs *(Dia setmana, Mes de any, Dia de mes, Hora de dia, Desc. tipus vehicle implicat, Descripció sexe, Edat, Descripció tipus persona, Descripció victimització)* i 11.844 observacions. Finalment, procedim a crear el fitxer "Accidents_BCN_2016.csv" que recull el conjunt de dades abans mencionat.

```
write.csv(AccidentsOK, "Accidents_BCN_2016.csv")
```

Figura 5. Creació del nou fitxer del conjunt de dades final.

4. Anàlisi de les dades.

1. Selecció dels grups de dades.

D'entrada, se seleccionen els grups de dades del nostre conjunt que resulten interessants per ser analitzats i aquests conjunts es presenten a la figura 6. S'ha volgut elegir un gran ventall de possibilitats d'estudi però també és cert, que no tots s'utilitzaran en les proves estadístiques.

```
#Agrupem les víctimes segons sexe
AccidentsOK.homes <- AccidentsOK[AccidentsOK$`Descripció sexe` == 0 ,]
AccidentsOK.dones <- AccidentsOK[AccidentsOK$`Descripció sexe` == 1 ,]

#Agrupem per tipus de vehicle implicat
AccidentsOK.bus <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 0 ,]
AccidentsOK.busArt <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 1 ,]
AccidentsOK.autocar <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 2 ,]
AccidentsOK.bici <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 3 ,]
AccidentsOK.cam <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 4 ,]
AccidentsOK.camSup <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 5 ,]
AccidentsOK.ciclo <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 6 ,]
AccidentsOK.cuadri <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 7 ,]
AccidentsOK.cuadriSup <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 8 ,]
AccidentsOK.furgo <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 9 ,]
AccidentsOK.maqui <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 10 ,]
AccidentsOK.moto <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 11 ,]
AccidentsOK.otro <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 12 ,]
AccidentsOK.taxi <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 13 ,]
AccidentsOK.4x4 <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 14 ,]
AccidentsOK.tracto <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 15 ,]
AccidentsOK.tren <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 16 ,]
AccidentsOK.turisme <- AccidentsOK[AccidentsOK$`Desc. Tipus vehicle implicat` == 17 ,]

#Agrupem pel resultat de l'accident
AccidentsOK.lleu <- AccidentsOK[AccidentsOK$`Descripció victimització` == 0 , ]
AccidentsOK.greu <- AccidentsOK[AccidentsOK$`Descripció victimització` == 1 , ]
AccidentsOK.mort <- AccidentsOK[AccidentsOK$`Descripció victimització` == 2 , ]
```

Figura 6. Grups de dades susceptibles d'anàlisi.

2. Comprovació de la normalitat i homogeneïtat de la variància.

Per la comprovació de la normalitat s'ha utilitzat la llibreria "nortest", prèviament instal·lada. Aquesta llibreria permet l'ús de la funció 'ad.test()' que permet analitzar la normalitat de la mostra segons la teoria d'*Anderson-Darling.*

```
        Anderson-Darling normality test

data:  as.integer(AccidentsOK$`Dia setmana`)
A = 234.02, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Mes de any`)
A = 191.57, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Dia de mes`)
A = 125.71, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Hora de dia`)
A = 73.165, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Desc. Tipus vehicle implicat`)
A = 792.17, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Descripció sexe`)
A = 2255.7, p-value < 2.2e-16

data:  as.integer(AccidentsOK$Edat)
A = 117.74, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Descripció tipus persona`)
A = 2111.1, p-value < 2.2e-16

data:  as.integer(AccidentsOK$`Descripció victimització`)
A = 4447.9, p-value < 2.2e-16
```

Figura 7. Resultats del test de normalitat.

S'observa que en tots els casos un valor de 'p-value' inferior a 0.05, per tant no es rebutja la hipòtesi nul·la i es conclou que les dades no segueixen una distribució normal.

Respecte a l'homogeneïtat de la variància, aquesta s'ha comprovat mitjançant la funció 'fligner.test()'. Aquesta funció utilitza la teoria de Fligner-Killeen per dur a terme la comprovació. Concretament, s'ha realitzat l'estudi de l'homogeneïtat fent referència al sexe de l'implicat a l'accident.

```
#Test homogeneitat de la variança
fligner.test(as.numeric(`Descripció victimització`) ~ as.numeric(`Descripció sexe`) , data= AccidentsOK)
```

Figura 8. Test de l'homogeneïtat de la variància.

El test parteix d'una hipòtesi que ambdues variàncies són nul·les, la figura 9 mostra els resultats del test.

```
        Fligner-Killeen test of homogeneity of variances

data:  as.numeric(`Descripció victimització`) by as.numeric(`Descripció sexe`)
Fligner-Killeen:med chi-squared = 3.0689, df = 1, p-value = 0.0798
```

Figura 9. Resultats del test de Fligner-Killeen.

Es pot observar que el resultat obtingut 'p-value' és superior a 0,05, per tant acceptem la hipòtesi que, en aquest cas, les variàncies són homogènies.

3. Proves estadístiques.

- PROVA 1: Contrast d'hipòtesi

La primera prova estadística que s'aplica consisteix en un contrast d'hipòtesi. Aquesta intenta donar resposta a la pregunta següent:

El sexe de la víctima influeix respecte al resultat de l'accident?

Per respondre a la pregunta, es creen dues mostres on la primera fa referència als accidents patits pels homes i la segona fa referència als accidents patits per dones.

```
accident.home.victimitzacio <-
  AccidentsOK[AccidentsOK$`Descripció sexe` == 0,]$`Descripció victimització`
accident.dona.victimitzacio <-
  AccidentsOK[AccidentsOK$`Descripció sexe` == 1,]$`Descripció victimització`
```

Figura 10. Mostres del test de contrast d'hipòtesi.

En aquest sentit, es procedeix amb el plantejament del contrast d'hipòtesi de dues mostres i es formula la següent hipòtesi alternativa:

$$\begin{cases} H_0: \mu_0 - \mu_1 = 0 \\ H_1: \mu_0 - \mu_1 > 0 \end{cases}$$

On:

- μ_0 fa referència a la mitja de la població que s'extreu de la primera mostra i μ_1 fa referència a la mitja que s'extreu de la segona.
- El valor del paràmetre $\alpha = 0{,}05$

```
t.test(as.integer(accident.home.victimitzacio), as.integer(accident.dona.victimitzacio),
       alternative = "t")
```

Figura 11. Formula del test.

```
        Welch Two Sample t-test

data:  as.integer(accident.home.victimitzacio) and as.integer(accident.dona.victimitzacio)
t = 1.5124, df = 10468, p-value = 0.1305
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.001310784  0.010164933
sample estimates:
 mean of x  mean of y
0.02242965 0.01800257
```

Figura 12. Resultats del test de contrast d'hipotesi.

S'obté un valor de "p-value" superior a 0,05. Per tant es rebutja la hipòtesi nul·la i s'arriba a la conclusió que el sexe de la víctima influeix en el resultat de l'accident.

- PROVA 2: Anàlisi de correlació

En segon lloc, és dur a terme un anàlisi de correlació entre variables a fi de definir quines variables influeixen més a l'hora de quantificar el dany que pateix la víctima en un accident. S'ha utilitzat el coeficient de correlació de *Kendall* que, iqual que el coeficient de Sperman, permet treballar amb dades que no segueixen una distribució normal. La figura 13 mostra el codi utilitzat per dur a terme els càlculs dels coeficients. A causa d'un problema de codi, no s'ha pogut automatitzar el codi i s'ha fet la matriu de coeficients manualment.

```
#Prova estadística: anàlisi de correlació
  #Creem la matriu de correlacions
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

  #Calculem el coeficient de correlació:
spearman_test = cor.test(as.integer(AccidentsOK$`Descripció tipus persona`),
                         as.integer(AccidentsOK$`Descripció victimització`),
                         method = "kendall")

{ corr_coef = spearman_test$estimate
p_val = spearman_test$p.value
# Add row to matrix
pair = matrix(ncol = 2, nrow = 1)
pair[1][1] = corr_coef
pair[2][1] = p_val
corr_matrix <- rbind(corr_matrix, pair)
rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(AccidentsOK)[8]
}

print(corr_matrix)
```

Figura 13. Codi pel càlcul dels coeficients de correlació mitjançant la teoria de Kendall.

	estimate	p-value
Dia setmana	0.009096963	2.573853e-01
Mes de any	-0.004733977	5.440816e-01
Dia de mes	0.014624700	5.484604e-02
Hora de dia	-0.005372882	4.851312e-01
Desc. Tipus vehicle implicat	-0.019473647	2.110756e-02
Descripció sexe	-0.017296472	5.953322e-02
Edat	0.042449173	2.031723e-08
Descripció tipus persona	0.050808992	1.097367e-08

Figura 14. Resultats de l'assaig.

Els resultats estan compresos entre els valors *-1* i *+1*, sent *-1* el valor per observacions totalment diferents i *+1* per les observacions totalment afins. Podem dir que, el fet de ser conductor, passatger o bé vianant és l'atribut que més correlació obté a l'hora de quantificar el dany que pateix la víctima en un accident, seguit per l'atribut 'edat'.

- PROVA 3: Model de regressió lineal.

En tercer i darrer lloc, és dur a terme un model de regressió lineal amb motiu de determinar els factors que influeixen més pel que fa al tipus de persona implicada en l'accident: conductor, passatger, vianant. Per això, s'avaluaran 3 models i s'elegirà el model que més s'adapti a la realitat.

```
#Prova estadística: model de regresió lineal
  #Variables:
diasem <- AccidentsOK$`Dia setmana`
mes <- AccidentsOK$`Mes de any`
diames <- AccidentsOK$`Dia de mes`
hora <- AccidentsOK$`Hora de dia`
vehicle <- AccidentsOK$`Desc. Tipus vehicle implicat`
sexe <- AccidentsOK$`Descripció sexe`
edat <- AccidentsOK$Edat
implicat <- AccidentsOK$`Descripció tipus persona`

resultat <- AccidentsOK$`Descripció victimització`

  #Models estudiats
{modelo1 <- lm(implicat ~  hora + sexe + vehicle, data=AccidentsOK)
  modelo2 <- lm(implicat ~  edat + sexe + vehicle, data=AccidentsOK)
  modelo3 <- lm(implicat ~  edat + diasem + vehicle, data=AccidentsOK)

  #Coeficients R^2 de cada model:
print('Modelo1:')
print(summary(modelo1)$r.squared)
print('Modelo2:')
print(summary(modelo2)$r.squared)
print('Modelo3:')
print(summary(modelo3)$r.squared)
}
```

Figura 15. Codi de la prova.

```
[1] "Modelo1:"
[1] 0.2274223
[1] "Modelo2:"
[1] 0.3518566
[1] "Modelo3:"
[1] 0.3085973
```

Figura 16. Bondat d'ajustament (R^2)

Podem observar que els resultats obtinguts no són significatius, donat que s'ha obtingut un valor de bondat d'ajustament màxim del 35%. Per tant no es pot extreure un model de regressió lineal que representi la realitat (només al 35%) ni tampoc cap conclusió amb fonament.

5. Conclusions.

En primer lloc, la realització de l'estudi ha permès posar en pràctica tot el après en el llarg del bloc 3. S'ha buscat un dataset que permeti donar resposta a una qüestió inicial i, posteriorment, s'ha tractat per adequar-lo l'estudi. Aquesta adequació ha passat per una selecció de dades d'interès i una posterior neteja de les dades. D'aquesta manera, s'ha disposat d'un dataset òptim per dur a terme l'anàlisi de les dades que passa per la realització de tres proves estadístiques *(Contrast d'hipòtesis, anàlisis de correlació i model de regressió lineal).*

En segon lloc, els resultats obtinguts en els anàlisis han permès observar que el sexe de la víctima va relacionat amb el resultat de l'accident (*gravetat de ferides*). I també s'ha pogut veure a través de l'anàlisi de correlació, que els danys patits en un accident estan relacionats amb l'edat i la condició de la víctima *(passatger, vianant o conductor).* Finalment, de l'última prova estadística no s'ha pogut extreure cap conclusió i es planteja modificar el model de regressió donat les característiques de les dades.

En darrer lloc, s'ha vist que la realització d'un anàlisi de dades no és una tasca senzilla ni breu en el temps. Però s'ha arribat a donar resposta a la pregunta inicial que fa referència als factors que més influencien en un accident de trànsit.

6. Annex

1.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Dia setmana	Dl	Dm	Dc	Dj	Dv	Ds	Dg											
Desc. Tipus vehicle implicat	Autobús	Autobús articulado	Autocar	Bicicleta	Camión <= 3,5 Tm	Camión > 3,5 Tm	Ciclomotor	Cuadriciclo <75cc	Cuadriciclo >=75cc	Furgoneta	Maquinaria de obras	Motocicleta	Otros vehíc. a motor	Taxi	Todo terreno	Tractocamión	Tranvía o tren	Turismo
Descripció sexe	Home	Dona																
Descripció tipus persona	Conductor	Passatger	Vianant															
Descripció victimització	Ferit lleu	Ferit greu	Mort															