

# Practica 2: Limpieza y validación de los datos

José Luis Fernández Losada – jfernandezlosada

Diciembre 2018

Github: <https://github.com/josele73/Prac02-Limpieza-y-validacion-de-los-datos>

Índice:

- 1- Descripción del dataset
- 2- Integración y selección de los datos
- 3- Limpieza de datos
- 4- Análisis y representación de los datos
- 5- Conclusiones
- 6- Recursos

## 1 – Descripción del Dataset

Este conjunto de datos es una lista de personas heridas que han estado involucradas en un accidente en la ciudad de Barcelona (España) durante el año 2015. Esta información es administrada por la policía en la ciudad de Barcelona.

El estudio de estos datos puede significar un mayor conocimiento de los accidentes en el casco urbano de Barcelona y su compresión aumentará la seguridad vial de la ciudad.

Se van a estudiar un fichero “2015\_accidents.csv” que contiene cabecera, 11780 registros y 25 variables.

El fichero contiene las siguientes variables:

- Número d'expedient: Número de expediente
- Codi districte: Código del distrito de Barcelona donde ha ocurrido el accidente
- Nom districte: Nombre del distrito
- Codi barri: Código del barrio de Barcelona donde ha ocurrido el accidente
- Nom barri: Nombre del barrio
- Codi carrer: Código de la calle
- Nom carrer: Nombre de la calle
- Num postal caption: Numero de postal de la calle
- Descripció dia setmana: Dia de la semana en que ocurre el accidente
- Dia setmana: Abreviatura del día de la semana en que ocurre el accidente
- Descripció tipus dia: Descripción del día, festivo o laboral.
- NK Any: Año
- Mes de any: Numero del mes
- Nom mes: Nombre del mes
- Dia de mes: Dia del mes
- Hora de dia: Hora del accidente
- Descripció causa vianant: Descripción del accidente.
- Desc. Tipus vehicle implicat: Tipo de vehículos implicados
- Descripció sexe: Sexo de la víctima
- Descripció tipus persona: Descripción del rol de la persona en el accidente.
- Descripció victimització: Descripción de la gravedad de los heridos.
- Coordenada UTM (Y): UTM coordenada Y
- Coordenada UTM (X): UTM coordenada X

```
> #Variables
> names(anyo2015)
[1] "Número.d.expedient"      "codi.districte"        "Nom.districte"       "codi.barri"
[5] "Nom.barri"              "codi.carrer"          "Nom.carrer"          "Num.postal.caption"
[9] "Descripció.dia.setmana"  "Dia.setmana"          "Descripció.tipus.dia" "NK.Any"
[13] "Mes.de.any"             "Nom.mes"              "Dia.de.mes"          "Hora.de.dia"
[17] "Descripció.causa.vianant" "Desc.Tipus.vehicle.implicat" "Descripció.sex"      "Descripció.tipus.persona"
[21] "Edat"                  "Descripció.victimització" "Coordenada.UTM..Y."   "Coordenada.UTM..X."
> |
```

Análisis de las variables:

### - Número d'expedient

Es un identificador del numero de expediente asignado por la policía al accidente. Existe un por cada herido.

En total hay 9104 expedientes distintos.

### - Codi.districte

Código del distrito en el que se produce al accidente. Asociado a la variable Nom.districte

```
> table(anyo2015$codi.districte)
 -1   1   2   3   4   5   6   7   8   9   10
11  612 3607 1240  822 1243  567  867  637  770 1404
> |
```

ç

### - Nom.districte

Descripción del distrito, asociado a la variable Codi.districte

```
> #Variable codigo distrito
> table(anyo2015$Nom.districte)

Ciutat Vella      Desconegut      Eixample      Gràcia      Horta-Guinardó      Les Corts
612                  11          3607        567          867          822
Nou Barris      Sant Andreu      Sant Martí      Sants-Montjuic      Sarrià-Sant Gervasi
637                  770          1404        1240          1243
```

> |

### - Codi.barri

Código del barrio en el que se produce al accidente. Asociado a la variable Nom.barri

---

```
> #Variable codigo barrio
> table(anyo2015$codi.barri)

-1   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
11  173 160 147 132 335 417 1455 629 504 267 268 293 117 57 167 113 74 151 382 253 187 32 217 157 232
26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
482 123 155 14  70  199 129 124 20  143 47  104 26  77  31  103 18  174 69  95  20  8  49  68  72  31
52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73
100 95  17  4   9  90  40  143 220 94  63  120 176 167 155 129 192 170 63  196 81  75
> |
```

## - Nom.barri

Descripción del nombre del barrio, asociado a la variable Codi.barri

> #variable nombre barrio				
> table(anyo2015\$Nom.barri)				
Baró de Viver	Can Baró	Can Peguera	Canyelles	
40	20	8	68	
Ciutat Meridiana	Desconeugut	Diagonal Mar i el Front Marítim del Poblenou	el Baix Guinardó	
4	11	170	124	
el Barri Gòtic	el Besòs i el Maresme	el Bon Pastor	el Camp d'en Grassot i Gràcia Nova	
160	63	143	129	
el camp de l'Arpa del clot	el Carmel	el Clot	el coll	
176	104	167	14	
el Congrés i els Indians	el Fort Pienc	el Guinardó	el Camp d'en Grassot i Gràcia Nova	
63	335	143	155	
el Poble Sec	el Poblenou	el Putxet i el Farró	el Raval	
268	192	123	173	
el Turo de la Peira	Horta	Hostafrancs	l'Antiga Esquerra de l'Eixample	
20	174	167	629	
la Barceloneta	la Bonanova	la Cova	la Dreta de l'Eixample	
147	113	18	1455	
la Font d'en Fargues	la Font de la Guatlla	la Guineueta	la Marina de Port	
47	57	49	117	
la Marina del Prat Vermell	la Maternitat i Sant Ramon	la Nova Esquerra de l'Eixample	la Prosperitat	
93	253	504	100	
la Sagrada Família	la Sagrera	la Salut	la Teixonera	
417	94	70	26	
la Trinitat Nova	la Trinitat Vella	la vall d'Hebron	la verneda i la Pau	
95	90	103	75	
Ta Vila de Gràcia	la vila olímpica del Poblenou	Tes Corral	Tes Roquetes	
109	129	182	72	
les Tres Torres	Montbau	Navas	Pedralbes	
157	31	120	187	
Porta	Provençals del Poblenou	Sant Andreu	Sant Antoni	
95	196	220	267	
Sant Genís dels Agudells	Sant Gervasi - Galvany	Sant Gervasi - la Bonanova	Sant Martí de Provençals	
77	482	232	81	
Sant Pere, Santa Caterina i la Ribera	Sants	Sants - Badal	Sarrà	
132	151	74	217	
Torre Baró	Vallbona	Vallcarca i els Penitents	vallvidrera, el Tibidabo i les Planes	
17	9	155	32	
verdun	69			
31				

## - Codi.carrer y Nom.carrer

Código de la calle en el que se produce al accidente. Asociado a la variable Nom.carrer

Hay 141 registros con valor “-1”

## - Num.postal.caption

Número postal de la dirección, relacionado con Num.carrer

## - Descripcio.dia.setmana

La columna contiene valores de los días de la semana.

```
> table(anyo2015$Descripció.dia.setmana)
```

Dijous	Dilluns	Dimarts	Dimecres	Dissabte	Diumenge	Divendres
1961	1662	1826	1934	1308	1033	2056

## - Dia.setmana

La columna contiene valores de los días de la semana abreviados, está relacionado con columna Descripcio.dia.setmana

```
> #Dia de la semana
> table(anyo2015$Dia.setmana)
```

DC	Dg	Dj	Dl	Dm	Ds	Dv
1934	1033	1961	1662	1826	1308	2056

### -Descripcio.tipus.dia

Todos los datos de esta variable tienen el valor “laboral”.

Estos datos están sesgados, no es posible que no existan accidentes en festivo.

Esta variable se **descartará**.

```
> table(anyo2015$Descripció.tipus.dia)
```

```
Laboral  
11780
```

### -NK.any

Todos los datos son del año 2015. Este dato se **descarta**, podría ser de utilidad en un futuro si se compara con los datos de otros años.

```
> #Descripción NK.anyo  
> table(anyo2015$NK.Any)
```

```
2015  
11780  
> |
```

### -Mes.de.any

Número de accidentes por cada mes del año 2015. Los meses son valores numéricos en el rango [1:12]

```
> #Mes.de.any  
> table(anyo2015$Mes.de.any)  
  
 1   2   3   4   5   6   7   8   9   10  11  12  
934 895 1006 1008 1026 1023 1006 791 920 1031 1084 1056  
> |
```

### -Nom.de.mes

Número de accidentes por cada mes del año 2015, con el nombre completo del mes. Relacionado con Mes.de.any

```
> #Nom.mes  
> table(anyo2015$Nom.mes)  
  
Abril    Agost  Desembre  Febrer  Gener  Juliol  Juny  Maig  Març  Novembre  Octubre  Setembre  
1008     791    1056     895    934    1006   1023   1026   1006   1084    1031    920  
> |
```

### -Dia.de.mes

Total de heridos por cada día del mes. se encuentran en el rango [1-31]

```
> #Valores dia mes
> range(anyo2015$Dia.de.mes,na.rm=TRUE)
[1] 1 31
> table(anyo2015$Dia.de.mes)

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
282 393 400 414 381 378 343 398 451 407 363 334 409 361 405 436 421 365 433 397 386 406 418 373 375 357 389 399 336 372 198
> |
```

### -Hora.de.dia

Los valores de esta variable están en formato 12 horas y no son útiles para nuestro estudio. Debería estar en formato 24 horas.

Esta variable se **descarta**.

```
> range(anyo2015$Hora.de.dia,na.rm=TRUE)
[1] 1 12
> table(anyo2015$Hora.de.dia)

 1   2   3   4   5   6   7   8   9   10  11  12
1032 956 791 740 821 946 1126 1299 1166 956 935 1012
> |
```

### -Descripcio.causa.vianant

Descripción de la causa del accidente. Variable categórica con 5 posibles valores.

```
> table(anyo2015$Descripció.causa.vianant)

Altres creuar per fora pas de vianants      Desobeyir altres senyals Desobeyir el senyal del semàfor
                                              123                           271                               1
No és causa del vianant Transitar a peu per la calçada      11028                                40
                                              11028                                40
> |
```

### -Tipus.vehicle.implicit

Tipo de vehículos implicados en el accidente. Variable categorúica.

```
> #Tipo vehículo implicado
> table(anyo2015$Desc.,Tipus.vehicle.implicit)

Autobús    Autobús articulado        Autocar       Bicicleta     Camión <= 3,5 Tm     Camión > 3,5 Tm     Ciclomotor     Cuadriciclo <75cc
516          21                      3            633           39                  17             1067                5
Cuadriciclo >75cc   Furgoneta   Maquinaria de obras Microbus <=17 plazas Motocicleta Otros vehic. a motor Taxi Todo terreno
1              321                     3            2             5849                 17             398                 11
Tractocamión   Tranvía o tren   Turismo
6              6                      2865
> |
```

### -Descripción.sex

Variable sobre el sexo de las personas siniestradas.

Aparecen 3 valores como desconocidos, se descartarán posteriormente en los análisis que incluya esta variable.

```
> #Descripción sexo
> table(anyo2015$Descripció.sex)

Desconegut      Dona      Home
            3     4405    7372
> |
```

### -Descripcio.tipus.persona

Variable categórica, los resultados son correctos.

```
> #Descripción de los tipos de personas
> table(anyo2015$Descripció.tipus.persona)

Conductor Passatger   Vianant
      8202       2329      1249
> |
```

### -Edat

La variable Edat no es un numero entero como podíamos esperar porque aparece el valor “desconegut” 125 veces.

```
> #Valores edad
> table(anyo2015$Edat)

 0      1      10     101     106     11      12      13      14      15      16      17      18      19      2
24     13      26      1      29      30      41      39      43      98      158      168      229      237      283      306      324      350      340      311      344      286      323      338      325      298
20     21      22     23      24      25      26      27      28      29      30      31      30      31      32      31      30      31      32      31      30      31      32      33      34      35      36      37      38      39      40      2
229    237    283    306    324    350    340    311    344    286    25    323    338    325    298
34     35      36      37      38      39      40      41      42      43      44      45      46      47
320    292    291    277    282    296    25    246    261    247    205    238    231    191    170
48     49      50      51      52      53      54      55      56      57      58      59      60      61      62      63      64      65      66      67      68      69      70      71      72      73      74
189    184    32     184    176    170    139    148    115    141    114    92      103      99      78      75      69      58      44      41      47      53      45      55      40      33      34      25      32      16      31      24      34      27      37      25      27
61     62      63      64      65      66      67      68      69      70      71      72      73      74
78     75      69      58      44      41      47      53      45      21      55      40      33      34      25
75     76      77      78      79      80      81      82      83      84      85      86      87      88
32     16      31      24      34      27      27      37      25      27      20      10      15      22      10      9      6      4      1      1      1
8      9      90      91      92      94      96      101      106
> |
```

Eliminamos los valores desconocidos y convertimos la columna en numérico.

```
> #Comparamos la tabla con la anterior
> table(a2015OK$edat)

 0      1      2      3      4      5      6      7      8      9      10      11      12      13      14      15      16      17      18      19      20      21      22      23      24      25      26      27      28      29      30      31      32      33      34      35      36      37      38      39      40      41
24     13     16     25     25     32     23     21     27     22     26     29     30     30     41     39     43     98     158     168     229     237     283     306     324     350     340     311     344     286     323     338     325     298     320     292     291     277     282     296     246     261
42     43     44     45     45     46     47     48     49     50     51     52     53     54     55     56     57     58     59     60     61     62     63     64     65     66     67     68     69     70     71     72     73     74     75     76     77     78     79     80     81     82     83
247    205    238    231    191    170    189    184    184    176    170    139    148    115    141    114    92    103    99    78    75    69    58    44    41    47    53    45    55    40    33    34    25    32    16    31    24    34    27    37    25    27
84     85     86     87     88     89     90     91     92     94     96     101     106
20     10     15     20     15     8     10     9     6     4     1     1     1
> |
```

Calculamos la media, la mediana y los cuartiles

```
> #sumario del numerico  
> summary(a2015OK$Edat)  
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
 0.00   26.00   36.00   38.07   48.00 106.00  
> |
```

Y la desviación estándar de la muestra

```
> #Desviacion estandar  
> sd(a2015OK$Edat)  
[1] 16.02751  
> |
```

### -Descripció.victimització

Variable categórica con valores de la gravedad de los heridos  
[Leves,Graves,Muertos]

```
> #Descripcion victimas  
> table(anryo2015$Descripció.victimització)  
  
Ferit greu Ferit lleu      Mort  
197       11557          26  
> |
```

### -Coordenada.UTM..X y Coordenada.UTM..Y

Contienen valores extremos fuera de rango de las coordenadas UTM. Estos valores “-1” serán descartados en los análisis geoespaciales.

El resto de los valores pertenecen a las coordenadas UTM ERTS89 31N

## 2– Integración y selección de datos

### 2.1 – Selección de las variables.

Vamos a descartar para su análisis las columnas que están asociadas a otra columna que contiene su código de referencia.

Columnas descartadas:

- Numero.d.expediente: Es un contador ID con el número de expediente.
- Nom.districte: Se usará la columna “Codi.districte”
- Nom.barri: Se usará la columna “Codi.barri”
- Nom.carrer: Se usará la columna “Codi.carrer”
- Dia.setmana: Se utilizará la columna “Descripcio.dia.semmana” que incluye el nombre completo del día de la semana.
- NK.any: Conocemos que el año de los accidentes es el 2015. No aporta ningún valor si no se va a comparar con accidentes de otros años.
- Nom.de.mes: Se usará la columna mes.de.any

A partir de este momento vamos a trabajar con un data frame llamado a2015OK donde se eliminarán las columnas descartadas y las filas con valores desconocidos y en el que ya se han eliminado los heridos con edad “-1”.

## 2.2 Tipos de variables.

Estas son las 16 variables con las que vamos a trabajar

```
> str (a2015OK)
'data.frame': 11655 obs. of 16 variables:
 $ Codi.districte      : int -1 -1 10 10 10 10 10 10 10 ...
 $ Codi.barri           : int -1 64 64 64 64 64 64 64 ...
 $ Codi.carrer          : int -1 -1 224802 134801 95506 194406 194406 161407 297001 ...
 $ Num.postal.caption   : Factor w/ 1634 levels "0000_0000","000050000",...: 1634 1634 326 960 316 9 62 285 1047 1220 ...
 $ Descripció.dia.setmana: Factor w/ 7 levels "Dijous","Dilluns",...: 3 1 3 3 4 3 6 7 3 4 ...
 $ Mes.de.any            : int 8 6 12 4 11 3 2 2 9 4 ...
 $ Dia.de.mes             : int 4 25 22 7 25 17 1 20 29 8 ...
 $ Hora.de.dia            : int 4 7 2 5 2 12 4 5 2 10 ...
 $ Descripció.causa.vianant: Factor w/ 6 levels "Altres","Creuar per fora pas de vianants",...: 5 5 5 2 5 4 5 5 1 5 ...
 $ Desc..Tipus.vehicle.implicitat: Factor w/ 19 levels "Autobús","Autobús articulado",...: 13 19 13 13 13 7 13 19 13 13 ...
 $ Descripció.sexe         : Factor w/ 3 levels "Desconegut","Dona",...: 3 3 3 3 3 2 3 2 3 3 ...
 $ Descripció.tipus.persona: Factor w/ 3 levels "Conductor","Passatger",...: 1 1 1 3 1 3 2 1 1 1 ...
 $ Edat                   : num 36 54 49 13 49 68 33 37 49 23 ...
 $ Descripció.victimització: Factor w/ 3 levels "Ferit greu","Ferit lleu",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ Coordenada.UTM..Y.       : num -1 -1 4585145 4585448 4585215 ...
 $ Coordenada.UTM..X.       : num -1 -1 431991 431800 431576 ...
```

## 3 – Limpieza de datos

### 3.1 Búsqueda valores vacíos o nulos

Analizamos los datos de los ficheros en busca de valores NA.

```
> #Comprobamos los valores nulos
> summarise_all(a2015OK, funs(sum(is.na(.))))
```

	Codi.districte	Codi.barri	Codi.carrer	Num.postal.caption	Descripció.dia.setmana	Mes.de.any	Dia.de.mes	Hora.de.dia	Descripció.causa.vianant	Desc.Tipus.vehicle.implicat	Descripció.sexu	Descripció.tipus.persona	Edat	Descripció.victimització	Coordenada.UTM..Y.	Coordenada.UTM..X.	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
>																	

NO existen valores NA pero si hemos encontrado en los códigos de distritos, barrios y calles el valor “-1”.

Además, en los campos sexo aparece el valor “desconegut” desconocido.

Vamos a considerar los anteriores valores como no validos y los vamos a eliminar del dataset.

Los accidentes con coordenadas con valor “-1” no eliminan ahora ya que aportan información relevante en el dataset, se eliminarán en el estudio geoespacial de los accidentes.

Como resultado vamos a trabajar con el siguiente dataset.

```
> str (a2015OK)
'data.frame': 11511 obs. of 16 variables:
 $ Codi.districte      : int 10 10 10 10 10 10 10 10 10 ...
 $ Codi.barri           : int 64 64 64 64 64 64 64 64 64 ...
 $ Codi.carrer          : int 224802 134801 95506 194406 194406 194406 161407 297001 194406 297001 ...
 $ Num.postal.caption   : Factor w/ 1634 levels "0000 0000","000050000",...: 326 960 316 9 62 285 1047 1220 425 1223 ...
 $ Descripció.dia.setmana: Factor w/ 7 levels "Dijous","Dilluns",...: 3 3 4 3 6 7 3 4 1 4 ...
 $ Mes.de.any           : int 12 4 11 3 2 2 9 4 11 3 ...
 $ Dia.de.mes           : int 22 7 25 17 1 20 29 8 19 11 ...
 $ Hora.de.dia          : int 2 5 2 12 4 5 2 10 7 8 ...
 $ Descripció.causa.vianant: Factor w/ 6 levels "Altres","Creuar per fora pas de vianants",...: 5 2 5 4 5 5 1 5 5 5 ...
 $ Desc.Tipus.vehicle.implicat: Factor w/ 19 levels "Autobús","Autobús articulado",...: 13 13 13 7 13 19 13 13 13 13 ...
 $ Descripció.sexu       : Factor w/ 3 levels "Desconegut","Dona",...: 3 3 3 2 3 2 3 3 3 3 ...
 $ Descripció.tipus.persona: Factor w/ 3 levels "Conductor","Passatger",...: 1 3 1 3 2 1 1 1 1 1 ...
 $ Edat                  : num 49 13 49 68 33 37 49 23 31 31 ...
 $ Descripció.victimització: Factor w/ 3 levels "Ferit greu","Ferit lleu",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ Coordenada.UTM..Y.     : num 4585145 4585448 4585215 4585174 4585195 ...
 $ Coordenada.UTM..X.     : num 431991 431800 431576 431649 431644 ...
```

Un total de 16 variables y 11.511 registros frente a las 24 variables originales con 11.780 registros.

### 3.2 Identificación y tratamiento de valores extremos.

La mayoría de las variables del dataset son categóricas, pero podemos analizar la variable Edat en busca de valores extremos.

Voy a utilizar la función boxplots.stats para localizar los valores que parecen no congruentes con el resto del conjunto de datos de la edad.

```
> boxplot.stats(a20150K$Edat)
$stats
[1] 0 26 36 48 81

$n
[1] 11511

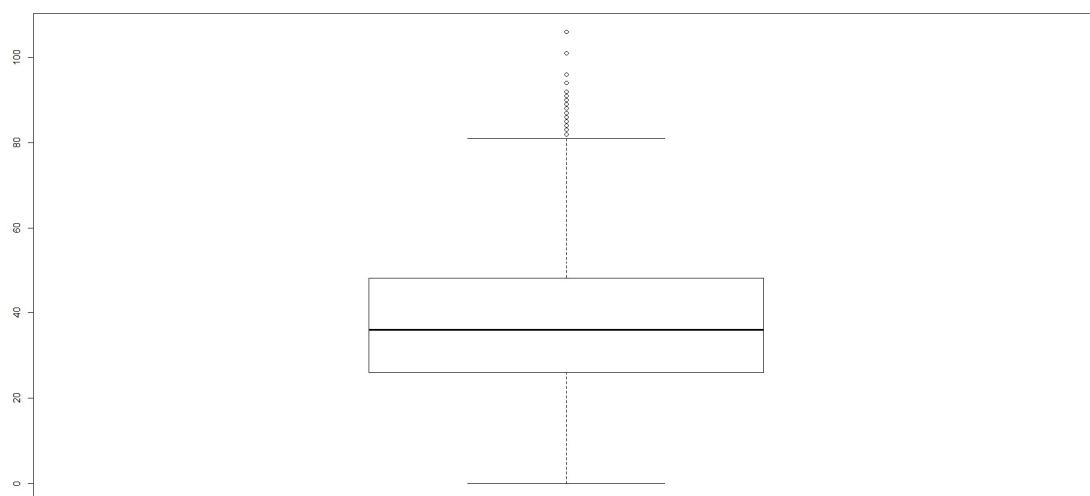
$conf
[1] 35.67602 36.32398

$out
[1] 89 94 85 82 92 88 87 89 88 90 94 86 82 87 83 89 86 96 86 85 84 87 86 84 83 84 87 87 85 91 88 87 106 86 83 91 82 87
[42] 84 83 83 87 88 85 87 90 83 82 92 92 84 84 89 94 88 92 84 82 86 82 86 92 92 101 88 91 84 87 90 90 88 85 86 82 82 91 89 89
[83] 82 82 87 88 83 83 82 83 83 84 87 84 84 83 90 82 85 82 84 87 83 89 87 87 84 82 82 83 87 86 83 82 84 86 83 84 83 90 83 91 84
[124] 83 87 88 88 83 85 90 83 82 86 88 84 82 89 90 91 86 83 91 90 83 88 87 83 83 91 82 94 85 88 82 82 90 84 85 84 86 82 83 86 83
[165] 82 88 83 91 86 82 87 88
> |
```

- \$stats: un vector de longitud 5, que contiene el extremo del inferior del bigote, la 'bisagra' inferior, la mediana, la 'bisagra' superior y el extremo del superior bigote.
- \$n: Numero total de elementos de la muestra sin contar NA.
- \$conf: Los extremos inferior y superior del notch
- \$out: Son los outliers, los valores de los puntos de datos que se encuentran más allá de los extremos de los bigotes.

La función nos detecta como valores extremos todos los que son superiores a 81, esto parece lógico teniendo en cuenta el histograma de la edad donde los valores con mas edad son escasos.

Aunque estos valores aparecen como extremos, son valores lógicos y no los voy a eliminar del Dataset, se trabajará con ellos.



## 4 – Análisis y representación de los resultados

Una vez que hemos seleccionado y limpiado los datos, vamos a proceder al análisis de estos.

El análisis de los accidentes se va a efectuar desde tres puntos de vista distinto:

- Análisis temporal
- Análisis de los heridos
- Análisis geoespacial de los accidentes.

### 4.1 Análisis temporal de los accidentes.

El objetivo de este análisis es encontrar un patrón en la relación de los accidentes y su ubicación en el tiempo.

Analizamos los días de la semana y los meses en que ocurren los accidentes.

#### Tabla de frecuencia cruzada meses y gravedad heridas

```
> table(a2015OK$Descripció.victimització,a2015OK$Nom.mes)

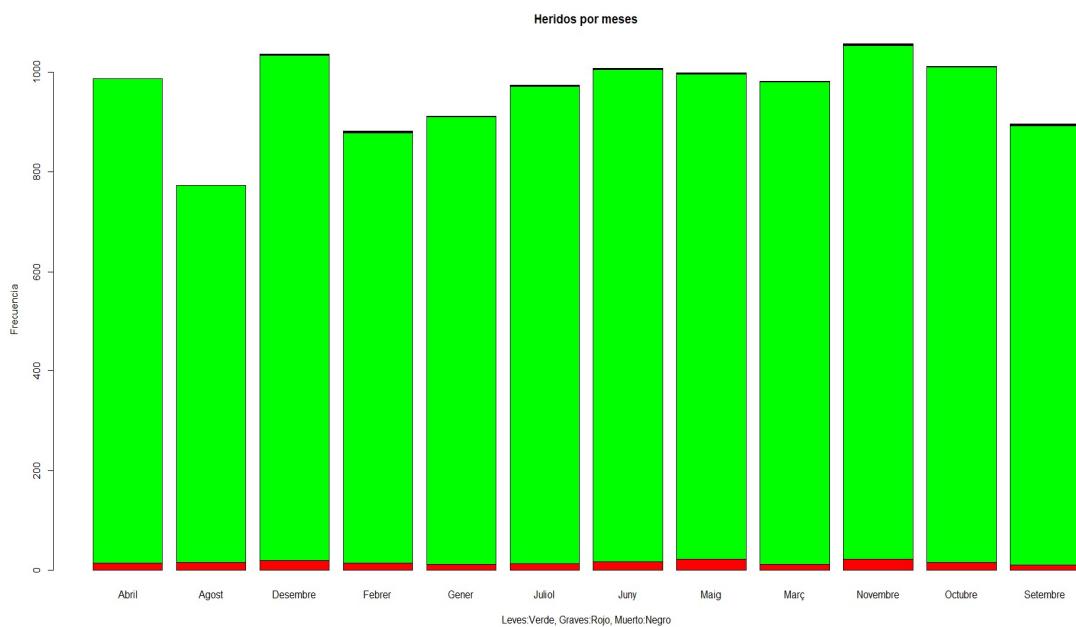
          Abril Agost Desembre Febrer Gener Juliol Juny Maig Març Novembre Octubre Setembre
Ferit greu  15    16     20    15    12    13    18    23    12    23    16    11
Ferit lleu  971   757   1013   863   898   958   986   973   968   1030   994   881
Mort       1      0      3      4      1      2      3      2      1      3      1      4
```

#### Probabilidad tabla de frecuencia cruzada meses y gravedad heridas

```
> prop.table(table(a2015OK$Descripció.victimització,a2015OK$Nom.mes),1)

          Abril Agost Desembre Febrer Gener Juliol Juny Maig Març Novembre Octubre Setembre
Ferit greu 0.07731959 0.08247423 0.10309278 0.07731959 0.06185567 0.06701031 0.09278351 0.11855670 0.06185567 0.11855670 0.08247423 0.05670103
Ferit lleu 0.08599008 0.06703861 0.08970953 0.07642579 0.07952533 0.08483882 0.08731846 0.08616720 0.08572441 0.09121502 0.08802692 0.07801984
Mort      0.04000000 0.00000000 0.12000000 0.16000000 0.04000000 0.08000000 0.12000000 0.08000000 0.04000000 0.12000000 0.04000000 0.16000000
```

## Heridos por meses



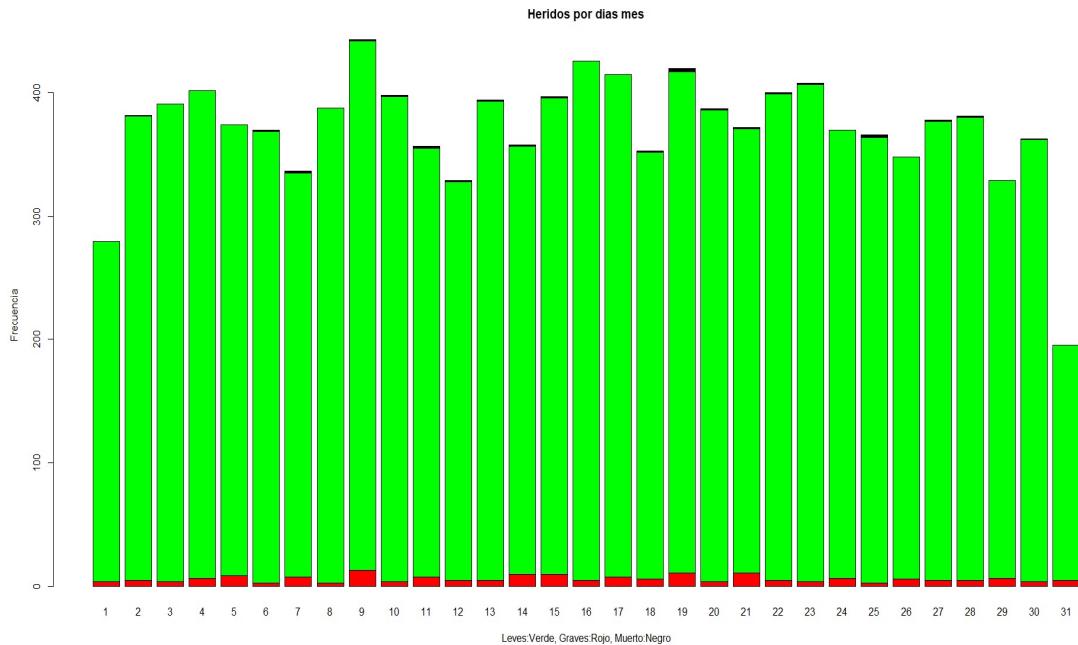
## Tabla de frecuencia cruzada días del mes y gravedad heridas

```
> table(a2015OK$Descripció.victimització,a2015OK$Dia.de.mes)
   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31
Ferit greu  4   5   4   7   9   3   8   3   13   4   8   5   5   10   10   5   8   6   11   4   11   5   4   7   3   6   5   7   4   5
Ferit lleu 276 376 387 395 365 366 327 385 429 393 347 323 388 347 386 421 407 346 406 382 360 394 403 363 361 342 372 375 322 358 190
Mort      0   1   0   0   0   1   2   0   1   1   2   1   1   1   1   0   0   1   3   1   1   1   0   2   0   1   1   0   1   0
> |
```

## Probabilidad tabla de frecuencia cruzada días del mes y gravedad heridas

```
> prop.table(table(a2015OK$Descripció.victimització,a2015OK$Dia.de.mes),1)
   1   2   3   4   5   6   7   8   9   10  11  12  13  14
Ferit greu 0.02061856 0.02577320 0.02061856 0.03608247 0.04639175 0.01546392 0.04123711 0.01546392 0.06701031 0.02061856 0.04123711 0.02577320 0.02577320 0.05154639
Ferit lleu 0.02444208 0.03329791 0.03427205 0.03498052 0.03232377 0.03241233 0.02895855 0.03409493 0.03799150 0.03480340 0.03072972 0.02860432 0.03436061 0.03072972
Mort      0.00000000 0.04000000 0.00000000 0.00000000 0.04000000 0.08000000 0.00000000 0.04000000 0.04000000 0.08000000 0.04000000 0.04000000 0.04000000 0.04000000
   15  16  17  18  19  20  21  22  23  24  25  26  27  28
Ferit greu 0.05154639 0.02577320 0.04123711 0.03092784 0.05670103 0.02061856 0.05670103 0.02577320 0.02061856 0.03608247 0.01546392 0.03092784 0.02577320 0.02577320
Ferit lleu 0.03418349 0.03728303 0.03604322 0.03064116 0.03595466 0.03382926 0.03188098 0.03489196 0.03568898 0.03214665 0.03196954 0.03028693 0.03294368 0.03320935
Mort      0.04000000 0.00000000 0.00000000 0.04000000 0.12000000 0.04000000 0.04000000 0.04000000 0.04000000 0.00000000 0.08000000 0.00000000 0.04000000 0.04000000
   29  30  31
Ferit greu 0.03608247 0.02061856 0.02577320
Ferit lleu 0.02851576 0.03170386 0.01682607
Mort      0.00000000 0.04000000 0.00000000
> |
```

## Heridos por día del mes



## Tabla de frecuencia cruzada días de la semana y gravedad heridas

```
> table(a2015OK$Descripció.victimització,a2015OK$Descripció.dia.setmana)

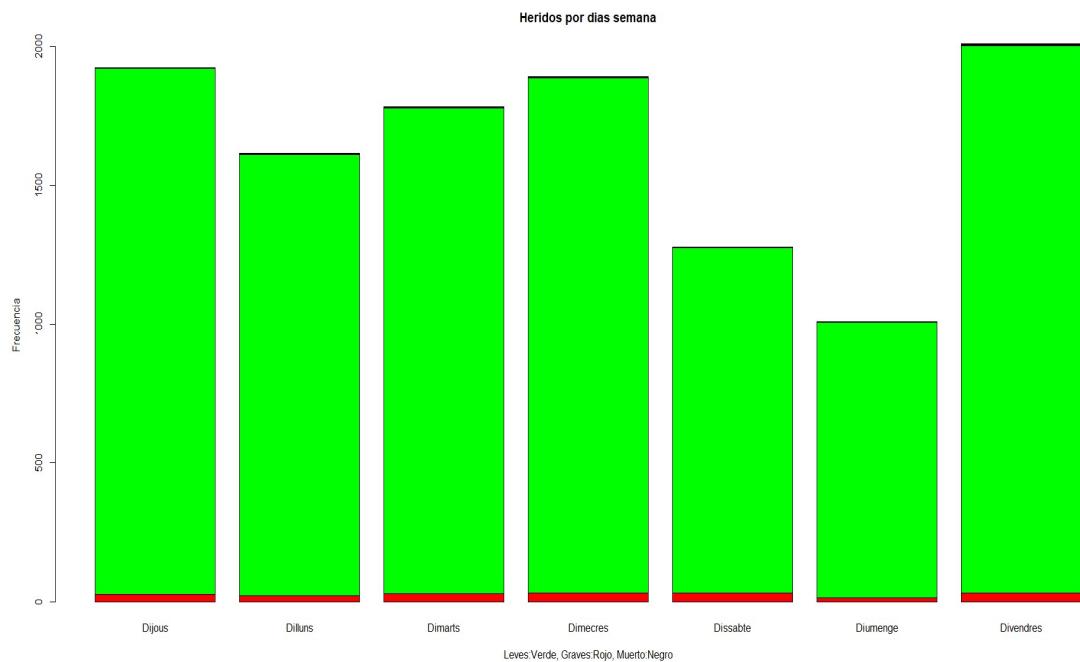
  Dijous Dilluns Dimarts Dimecres Dissabte Diumenge Divendres
Ferit greu   27      24     30     34     32     15     32
Ferit lleu  1894    1588   1749   1853   1244   993   1971
Mort         2       3      4      5      3      2      6
> |
```

## Probabilidad tabla de frecuencia cruzada días de la semana y gravedad heridas

```
> prop.table(table(a2015OK$Descripció.victimització,a2015OK$Descripció.dia.setmana),1)

  Dijous   Dilluns   Dimarts   Dimecres   Dissabte   Diumenge   Divendres
Ferit greu 0.13917526 0.12371134 0.15463918 0.17525773 0.16494845 0.07731959 0.16494845
Ferit lleu 0.16772937 0.14063053 0.15488842 0.16409848 0.11016649 0.08793836 0.17454835
Mort      0.08000000 0.12000000 0.16000000 0.20000000 0.12000000 0.08000000 0.24000000
> |
```

## Heridos por día de la semana



## 4.2 Análisis social de los heridos.

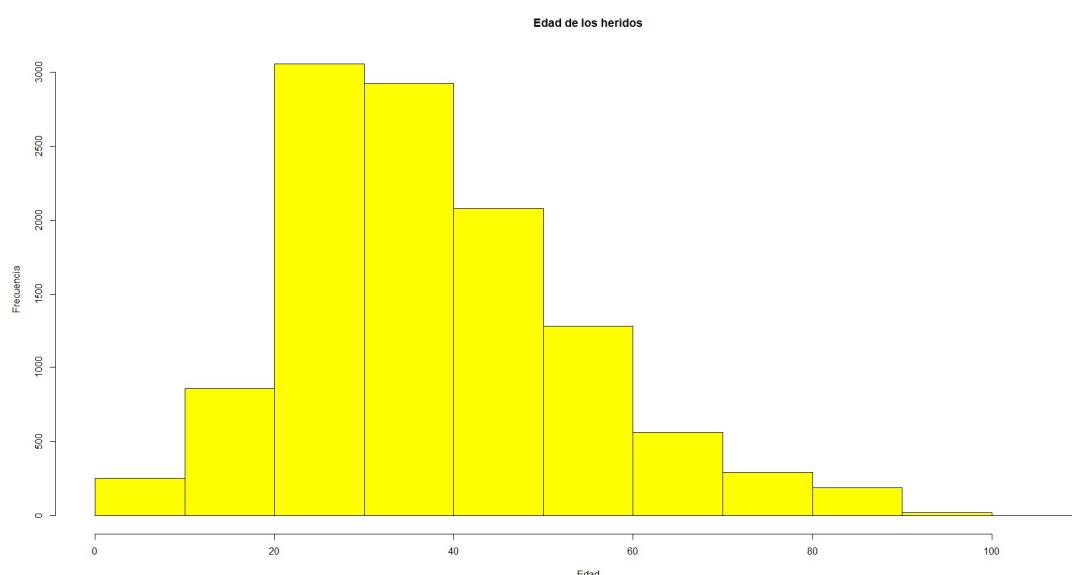
Antes de empezar el análisis por edad de los heridos vamos a eliminar los valores de edad desconocida.

En el año 2015 la edad mínima de los heridos en accidente de tráfico en Barcelona fue de 0 años y la máxima de 106.

La media es de 36 años y la mediana 48.

A continuación, el Histograma de las edades.

### Edad de los heridos



### Normalidad de la edad

Utilizando la prueba de normalidad Anderson-Darling, se considera que si obtenemos un valor por encima del valor 0.05 se considera que la variable sigue una distribución normal.

En nuestro caso el valor es inferior y por lo tanto la variable edad no sigue una distribución normal.

```
> (ad.test(a2015OK$Edat)$p.value)
[1] 3.7e-24
> if ((ad.test(a2015OK$Edat)$p.value)>0.05)
+ { print("Distribucion normal")
+ } else
+ {print("No es una distribucion normal")}
[1] "No es una distribucion normal"
> |
```

## Homogeneidad de la varianza de la edad y sexo

Usamos el test de Fligner-Killeen para estudiar la homogeneidad de la varianza de los grupos de edad para hombres y mujeres.

En nuestro estudio, el valor es superior al 0.05 y por lo tanto ambas muestras son homogéneas

```
> #Homogeneidad de las varianzas
> fligner.test(a2015OK$Edat~a2015OK$Descripción.sexu, data=a2015OK)

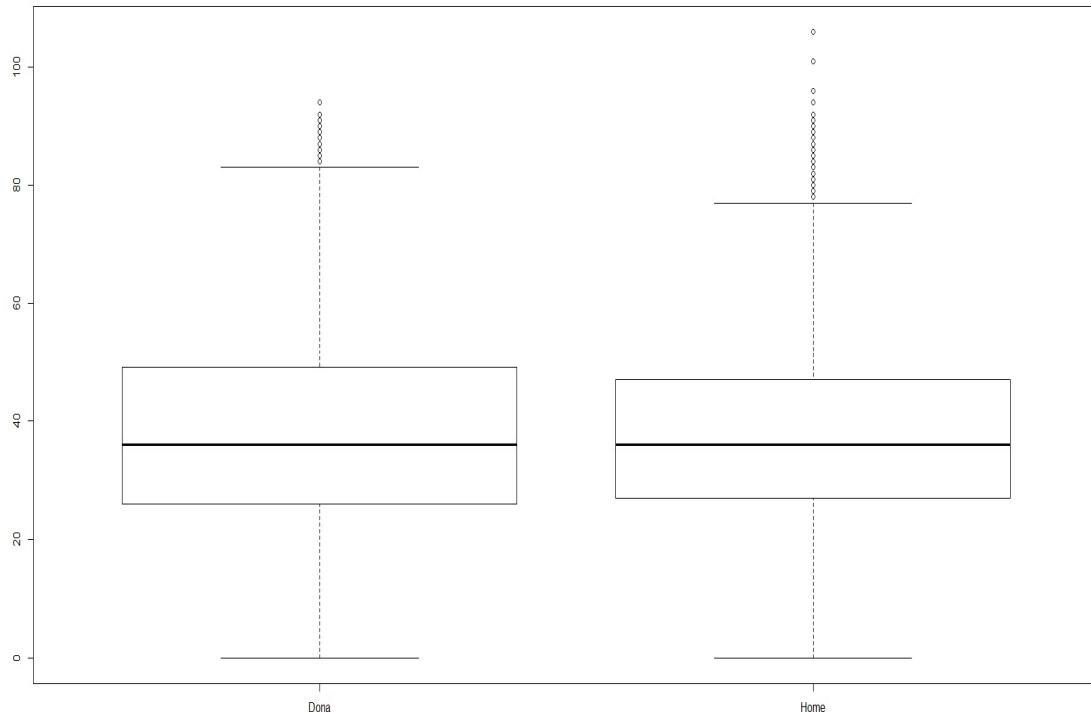
Fligner-Killeen test of homogeneity of variances

data: a2015OK$Edat by a2015OK$Descripción.sexu
Fligner-Killeen:med chi-squared = 93.369, df = 1, p-value < 2.2e-16

> if ((fligner.test(a2015OK$Edat~a2015OK$Descripción.sexu, data=a2015OK))>0.05)
+ { print("Muestras homogeneas")
+ }else
+ {print("Muestras no homogeneas")}
[1] "Muestras homogeneas"
```

## Caculo de la desviación edad por sexo

```
> aggregate(a2015OK$Edat,by=list(a2015OK$Descripción.sexu),mean,na.rm=TRUE)
  Group.1      x
1   Dona 39.02341
2   Home 37.52599
> |
```



## Frecuencias por edad - Sexo

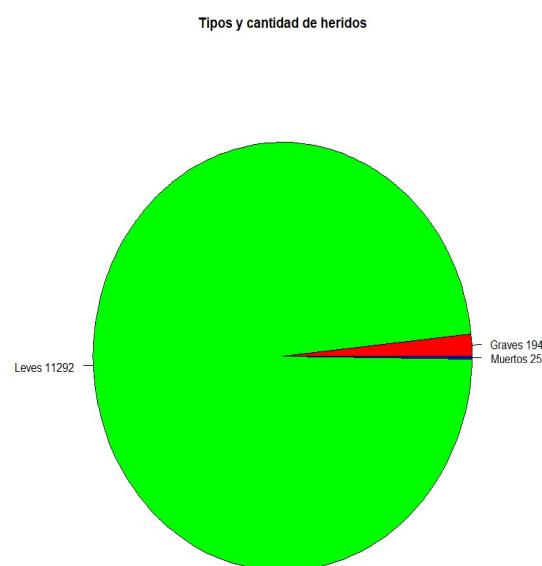
```
> summary(edadsexo)
   a2015OK.Edat      a2015OK.Descripció.sex
  Min.   : 0.00   Dona:4315
  1st Qu.: 26.00  Home:7196
  Median : 36.00
  Mean   : 38.09
  3rd Qu.: 48.00
  Max.   :106.00
> |
```

## Relación gravedad heridas - Sexo

```
> #Relacion heridas y sexo
> table(a2015OK$Descripció.victimització,a2015OK$Descripció.sex)

          Dona Home
Ferit greu  63  131
Ferit lleu 4250 7042
Mort        2   23
> |
```

## Tipos y cantidad de heridos

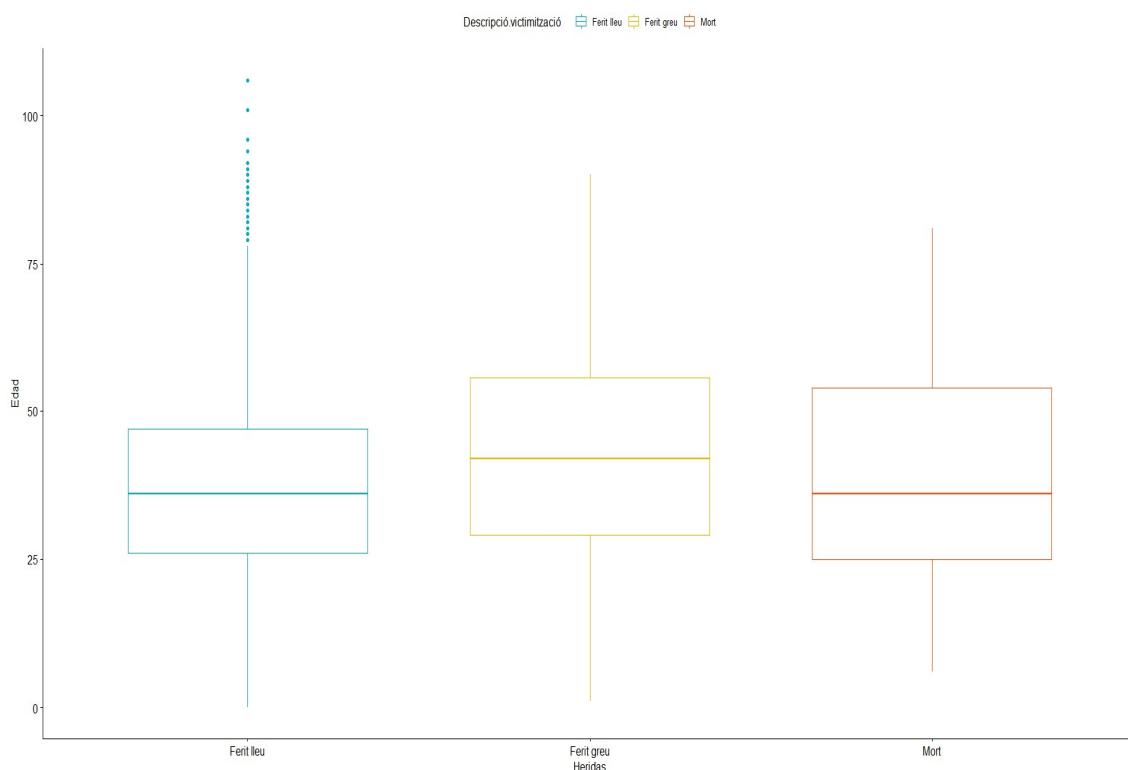


## Relación gravedad heridas - Edad

Mediante un test de Kruskal-Wallis, analizaremos si existen diferencias significativas entre las edades de los muertos, de los heridos leves y de los heridos graves

```
> group_by(a2015OK,Descripció.victimització) %>%
+   summarise(
+     count = n(),
+     mean = mean(Edat, na.rm = TRUE),
+     sd = sd(Edat, na.rm = TRUE),
+     median = median(Edat, na.rm = TRUE),
+     IQR = IQR(Edat, na.rm = TRUE)
+   )
# A tibble: 3 x 6
  Descripció.victimització count  mean    sd median   IQR
  <ord>          <int> <dbl> <dbl> <dbl> <dbl>
1 Ferit lleu        11292  38.0  16.0   36.  21.0
2 Ferit greu         194   43.6  19.9   42.  26.8
3 Mort                 25   39.8  18.7   36.  29.0
> |
```

Gráficamente:

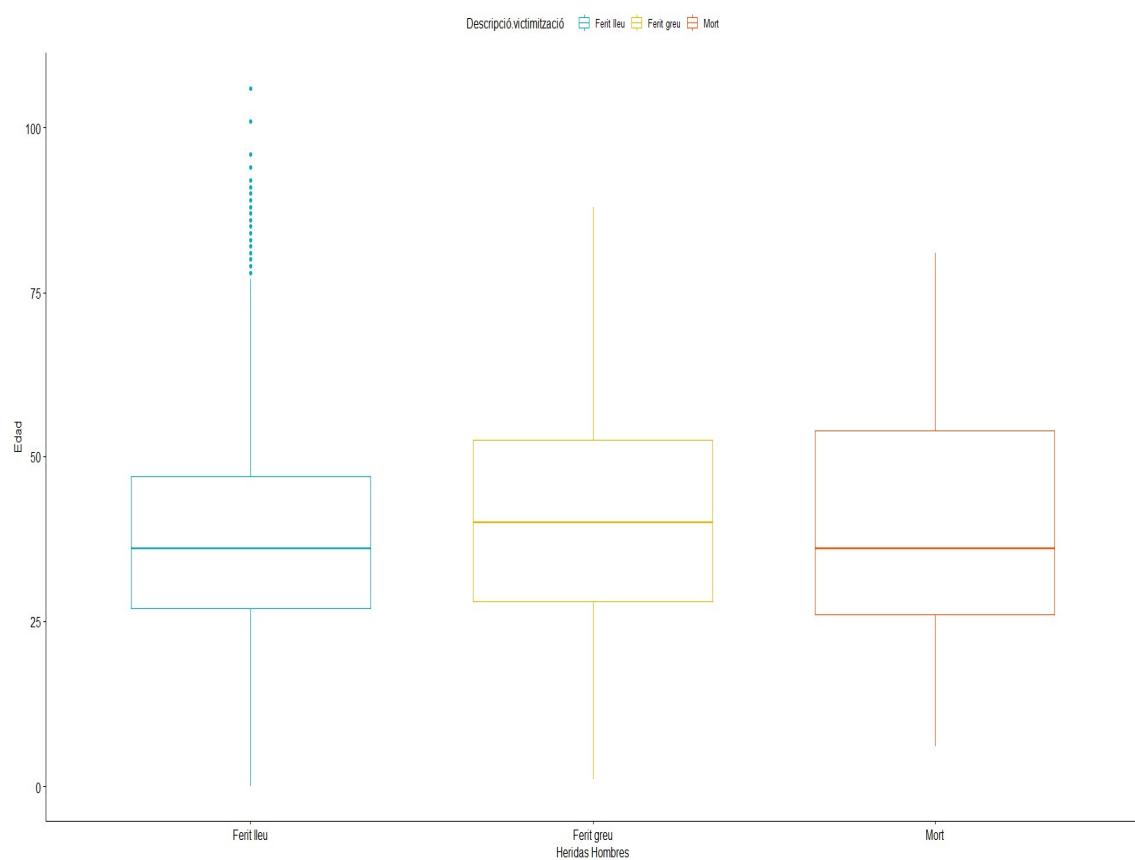


## Relación gravedad heridas – Hombres

Analizamos individualmente la muestra de los hombres

```
> group_by(Hombres,Descripció.victimització) %>%
+   summarise(
+     count = n(),
+     mean = mean(Edat, na.rm = TRUE),
+     sd = sd(Edat, na.rm = TRUE),
+     median = median(Edat, na.rm = TRUE),
+     IQR = IQR(Edat, na.rm = TRUE)
+   )
# A tibble: 3 x 6
  Descripció.victimització count  mean    sd median   IQR
  <fct>           <int> <dbl> <dbl> <dbl> <dbl>
1 Ferit lleu        131  42.5  19.5  40.  24.5
2 Ferit greu       2042  37.4  14.9  36.  20.0
3 Mort              23   40.0  18.7  36.  28.0
> |
```

Gráficamente el resultado es el siguiente:

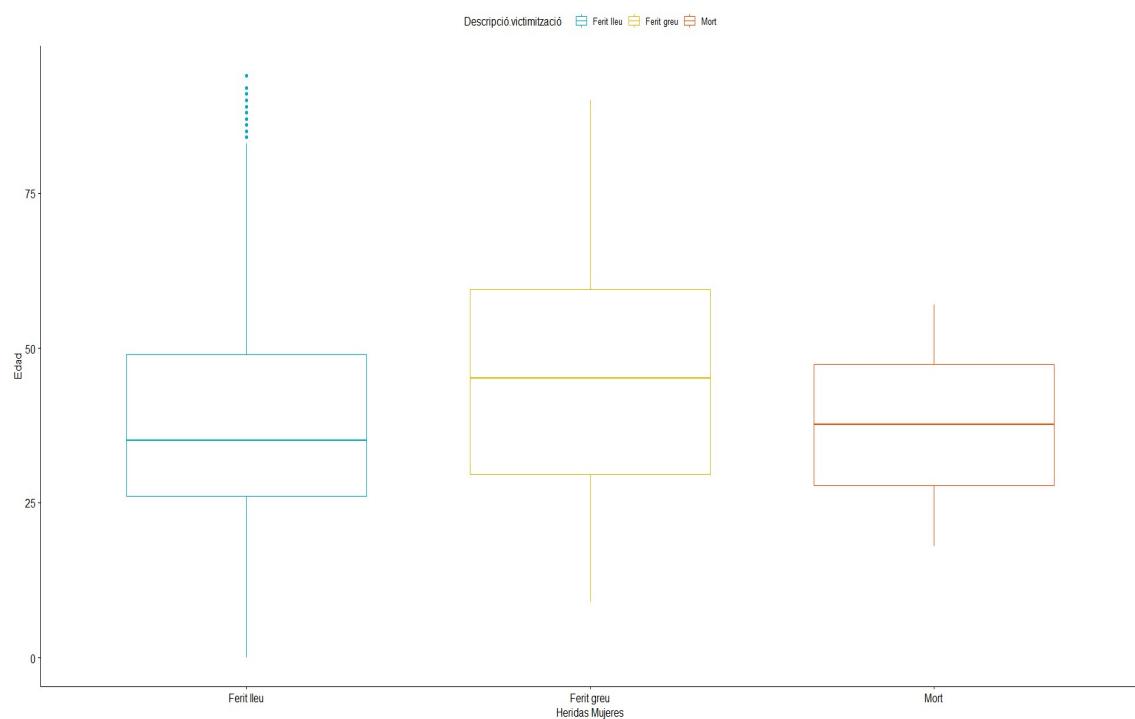


## Relación gravedad heridas – Mujeres

Analizamos individualmente la muestra de las mujeres

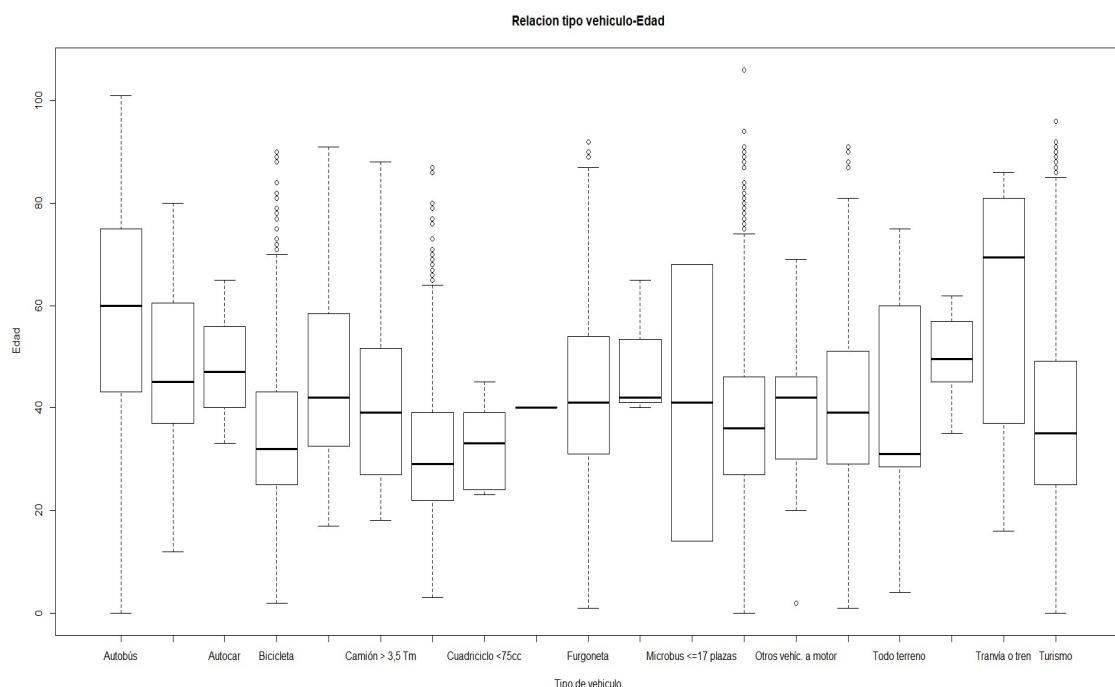
```
> group_by(Mujeres,Descripció.victimització) %>%
+   summarise(
+     count = n(),
+     mean = mean(Edat, na.rm = TRUE),
+     sd = sd(Edat, na.rm = TRUE),
+     median = median(Edat, na.rm = TRUE),
+     IQR = IQR(Edat, na.rm = TRUE)
+   )
# A tibble: 3 x 6
  Descripció.victimització count  mean    sd median   IQR
  <fct>           <int> <dbl> <dbl> <dbl> <dbl>
1 Ferit greu        63  45.9  20.8  45.0  30.0
2 Ferit lleu       4250  38.9  17.6  35.0  23.0
3 Mort              2  37.5  27.6  37.5  19.5
> |
```

Gráficamente el resultado es el siguiente:

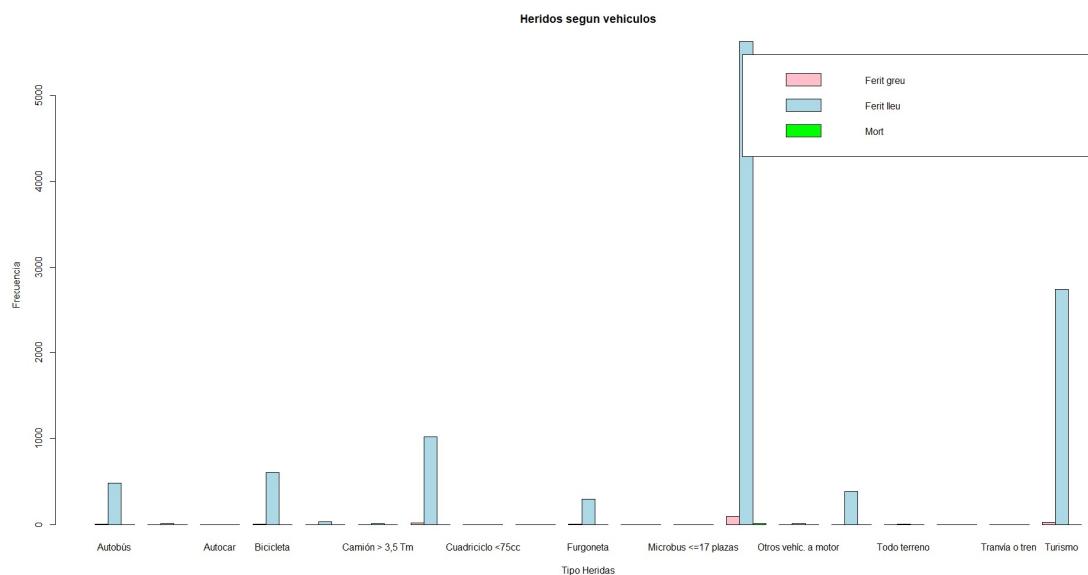


## Relación tipo de vehículo - Edad

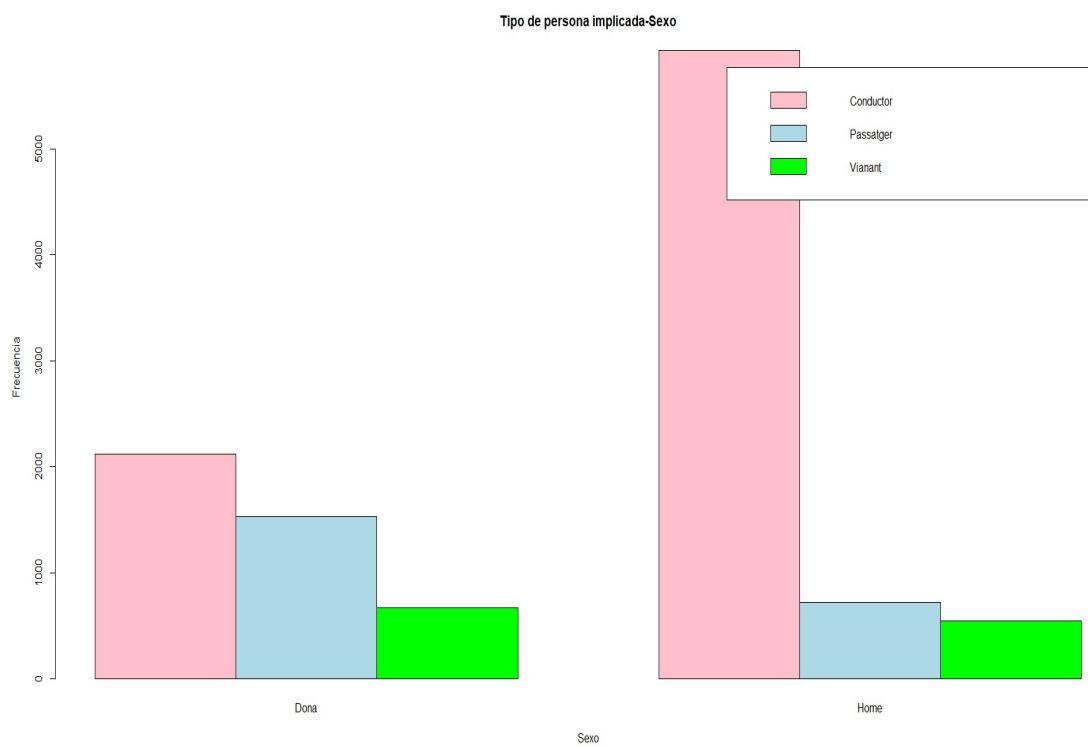
```
> group_by(a2015OK, Desc..Tipus.vehicle.implicit) %>%
+   summarise(
+     count = n(),
+     mean = mean(Edat, na.rm = TRUE),
+     sd = sd(Edat, na.rm = TRUE),
+     median = median(Edat, na.rm = TRUE),
+     IQR = IQR(Edat, na.rm = TRUE)
+   )
# A tibble: 19 x 6
  Desc..Tipus.vehicle.implicit count  mean      sd median    IQR
  <fct>                      <int> <dbl>    <dbl> <dbl>    <dbl>
1 Autobús                     498  57.8    21.2   60.0  31.8
2 Autobús articulado          19   47.5    17.6   45.0  23.5
3 Autocar                     3    48.3    16.0   47.0  16.0
4 Bicicleta                   621  35.0    15.7   32.0  18.0
5 Camión <= 3,5 Tm            39   47.3    19.0   42.0  26.0
6 Camión > 3,5 Tm             16   41.8    20.4   39.0  23.2
7 Ciclomotor                  1044 32.0    12.5   29.0  17.0
8 Cuadriciclo <75cc           5    32.8    9.50   33.0  15.0
9 Cuadriciclo >=75cc          1    40.0    NaN    40.0  0.
10 Furgoneta                   306  43.9    19.0   41.0  22.8
11 Maquinaria de obras          3   49.0    13.9   42.0  12.5
12 Microbus <=17 plazas        2    41.0    38.2   41.0  27.0
13 Motocicleta                  5743 37.3    12.9   36.0  19.0
14 Otros vehíc. a motor         17   39.2    16.3   42.0  16.0
15 Taxi                         391  40.7    16.2   39.0  22.0
16 Todo terreno                 11   40.3    22.6   31.0  31.5
17 Tractocamión                6    49.7    9.46   49.5  9.75
18 Tranvía o tren                6   59.8    27.4   69.5  33.8
19 Turismo                      2780 37.8    18.5   35.0  24.0
> |
```



## Relación tipo de vehículo - Gravedad heridos



## Relación tipo de persona implicada - Sexo

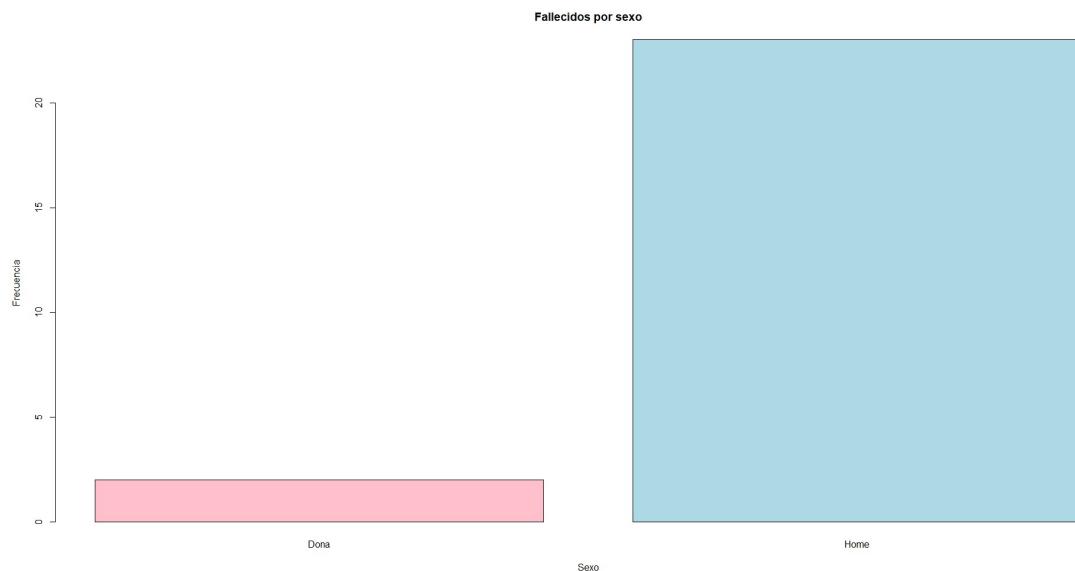


## 4.3 Análisis fallecidos

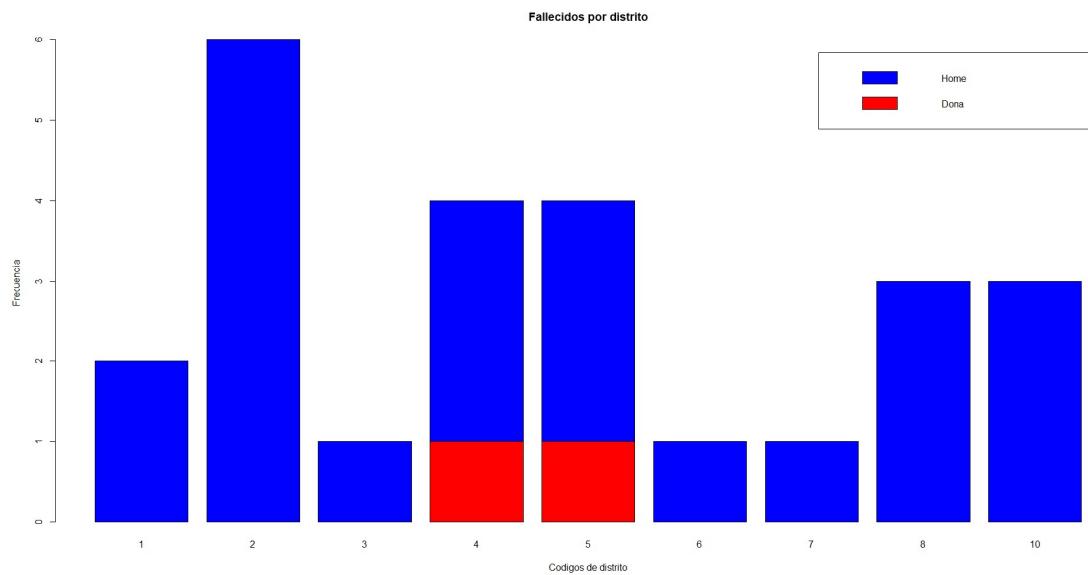
### Relación gravedad fallecidos - Edad

```
> group_by(Muertos,Descripció.sexu) %>%
+   summarise(
+     count = n(),
+     mean = mean(Edat, na.rm = TRUE),
+     sd = sd(Edat, na.rm = TRUE),
+     median = median(Edat, na.rm = TRUE),
+     IQR = IQR(Edat, na.rm = TRUE)
+   )
# A tibble: 2 x 6
  Descripció.sexu count  mean    sd median   IQR
  <fct>        <int> <dbl> <dbl> <dbl> <dbl>
1 Dona           2     37.5  27.6  37.5  19.5
2 Home          23    40.0  18.7  36.0  28.0
> |
```

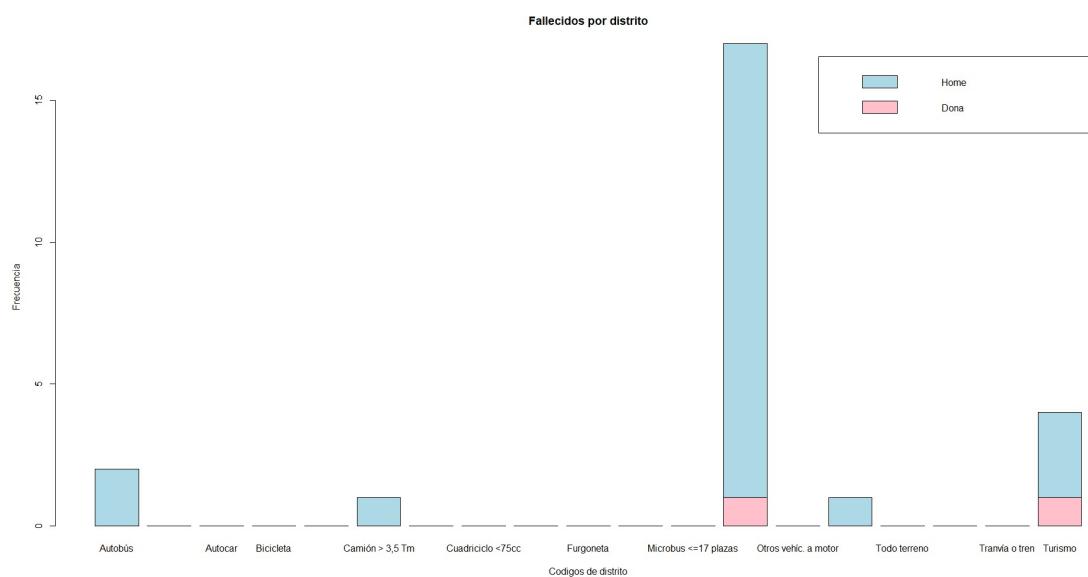
### Total fallecidos por sexo



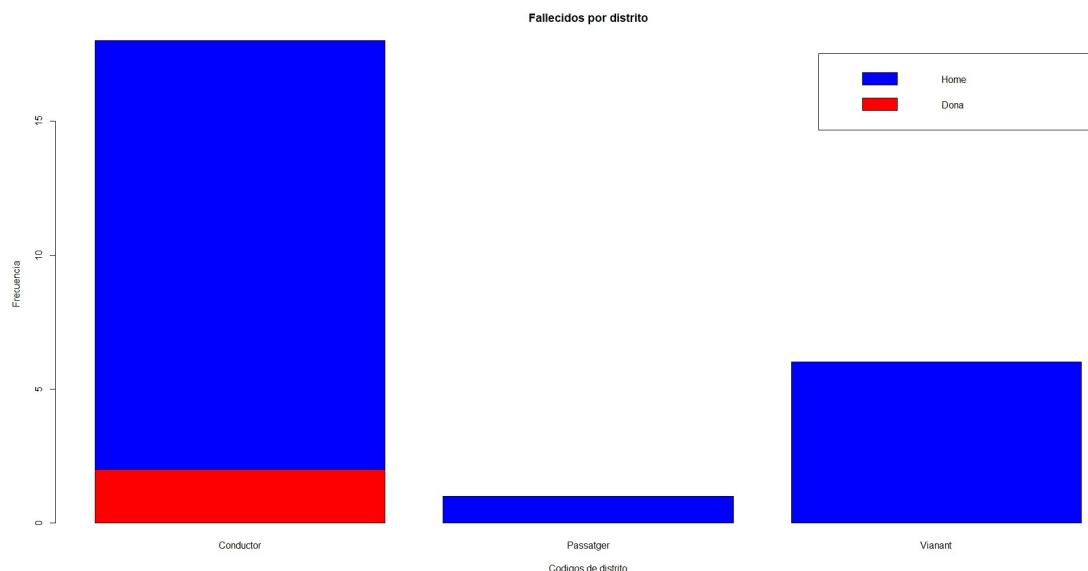
## Relación fallecidos Sexo - Barrio



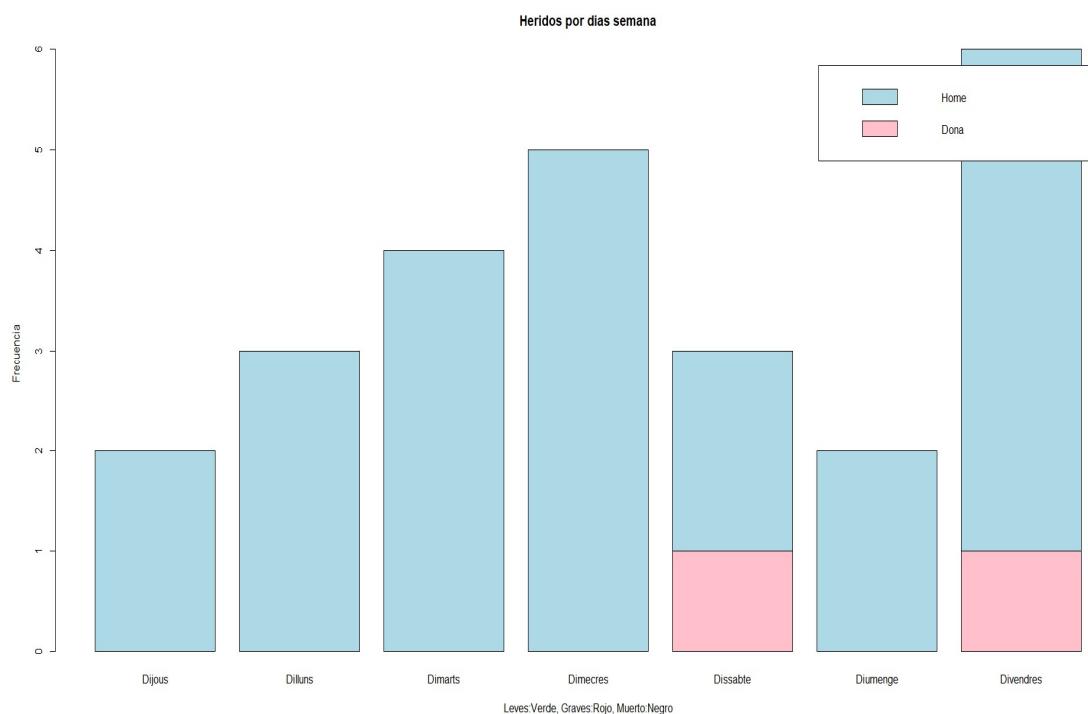
## Relación fallecidos – Tipo vehículo



## Relación fallecidos – Tipo persona involucrada



## Relación fallecidos – Dia de la semana



#### 4.4 Análisis geoespacial de los accidentes.

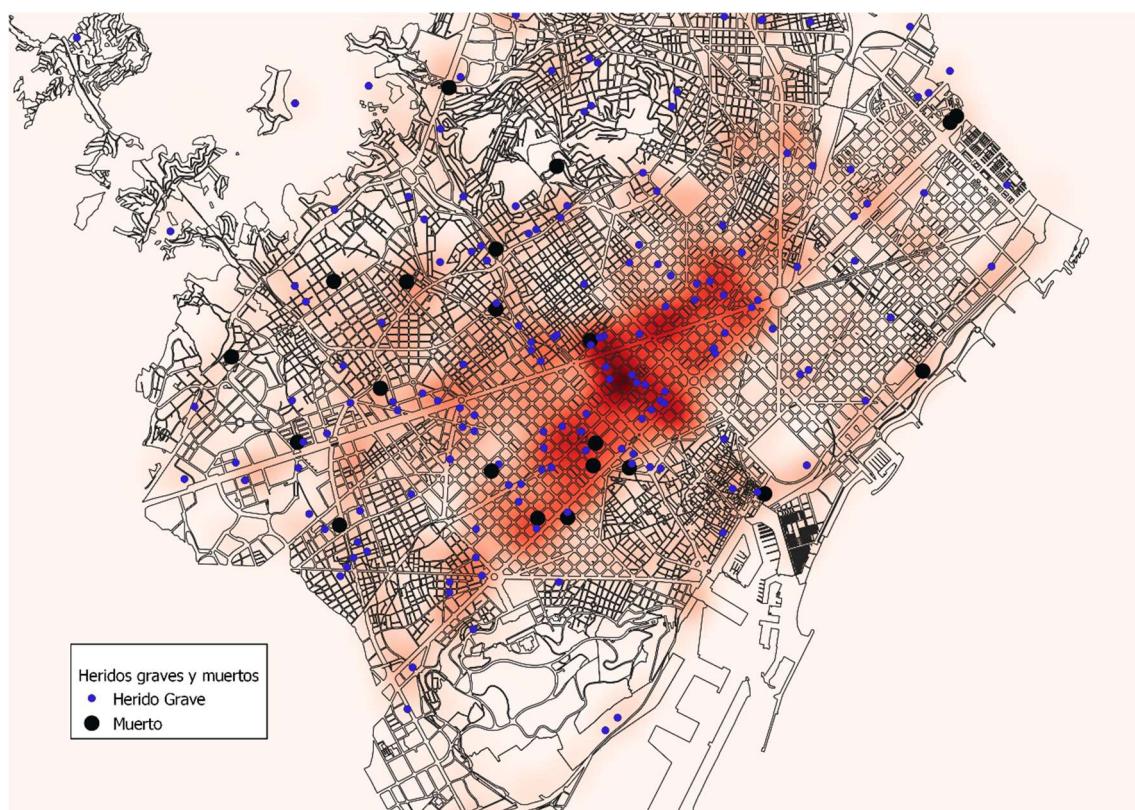
Antes de empezar el análisis vamos a eliminar los heridos no georreferenciados, con valores de '-1'. En total se eliminan 141 registros del dataset.

Utilizaremos la cartografía de la ciudad descargada desde la web de catastro.

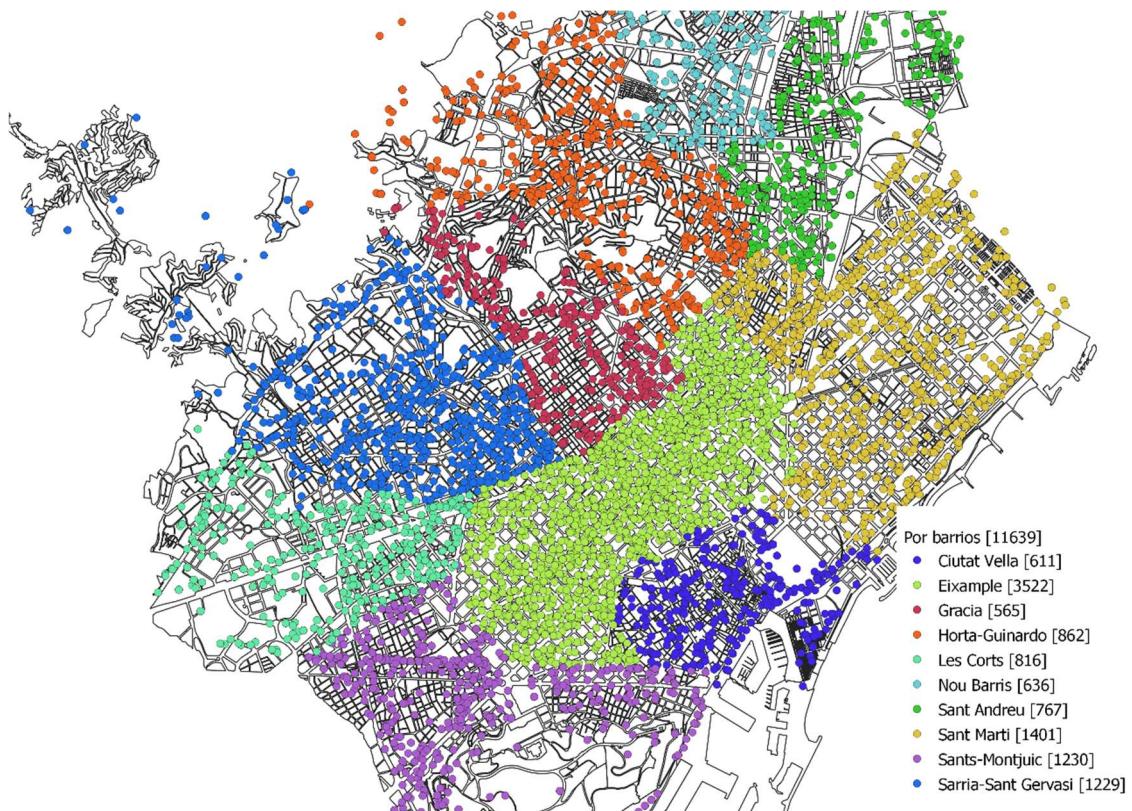
Aunque es posible hacer el análisis en R, visualmente no es muy efectivo y se va a efectuar sobre QGIS.

#### Mapa de calor de los accidentes sobre cartografía Barcelona

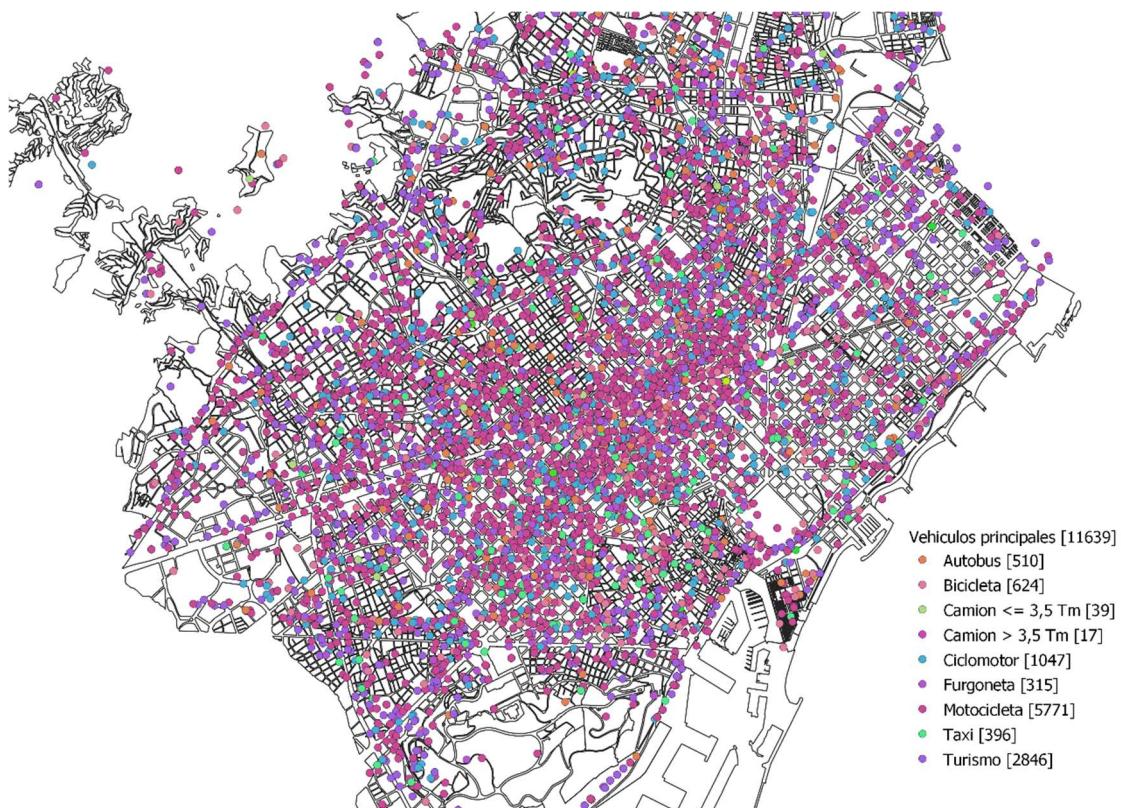
Mapa de calor con posicionamiento de los heridos graves y muertos



### Mapa de los accidentes por barrios



### Mapa de los accidentes según vehículos principales



## 5 – Conclusiones

En una ciudad como Barcelona, con mas de 1.600.000 habitantes, se producen en el año 2015 la cantidad de 11780 heridos en 9104 accidentes de tráfico con 26 fallecidos.

El perfil del herido en accidente es de un varón de algo menos de 40 años que es conductor de una motocicleta y que resulta herido leve en la zona de L’Eixample.

En cuanto al sexo de los heridos podemos afirmar que los hombres prácticamente duplican a las mujeres en número de heridos.

El perfil del fallecido es similar al del herido, es un varón de unos 40 años que conduce una motocicleta y que sufre el accidente en L’Eixample. Un dato significativo es que el 90% de los fallecidos son varones.

Los vehículos más implicados en accidentes son las motocicletas y según su rango de edad, los heridos que utilizan vehículos privados (vehículos de 2 ruedas y turismos) tienen una edad media inferior que los heridos en transporte publico tienen una edad superior a los anteriores.

En el estudio temporal se observa como los meses de noviembre y diciembre acumulan mas accidentes, el mes de agosto, suponemos que por las vacaciones y mejor tránsito por la ciudad, es el de menor siniestralidad.

El día de la semana más conflictivo es el viernes, tanto en número de accidentes con heridos como en fallecidos.

Como conclusión creemos necesaria una campaña de concienciación destinada principalmente a los conductores de motocicleta para disminuir su elevada siniestralidad.

Animamos a que se mejore la recolección de datos incluyendo el formato 24 horas y la distinción entre días festivos y laborables.

Por último, como mejora, se debe comparar este conjunto de datos con los datos de otros años para obtener tendencias a lo largo de los años.

## 6 – Recursos

- Fichero MASA.shp del catastro. Es necesario certificado digital para su descarga. <http://www.catastro.meh.es/>
- Fichero 2015\_accidents.csv <https://www.kaggle.com/marcvelmer/barcelona-accident> procedente de <http://opendata-ajuntament.barcelona.cat/en/>

**Alumno:**

**José Luis Fernández Losada**