

# Practica 2: Limpieza y validación de los datos

José Luis Fernández Losada – jfernandezlosada

Diciembre 2018

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 comprensión aumentará la seguridad vial de la ciudad.

## 2 – Integración y selección de datos

Se van a estudiar un fichero cvs distintos, cada uno de los cuales contiene la información sobre los accidentes el año indicado en el nombre del fichero:

- 2015\_accidents: Contiene Cabecera, 11780 registros y 25 variables.

Los dos ficheros contienen idénticas variables, son las siguientes:

- 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: Día 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: Día del mes
- Descripció torn: Descripción del momento del día del accidente
- 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

## 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 "Descripció.dia.setmana" 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:

Las columnas descartadas se extraerán junto a su código en el fichero "codificación.csv" para facilitar su uso posterior, si fuera necesario.

## 2.2 Tipos de variables.

Los tipos de las variables del dataset son los siguientes:

*str (anyo2015)*

```
> str (anyo2015)
'data.frame': 11780 obs. of 21 variables:
 $ Número.d.expedient : Factor w/ 9104 levels "2015S000001",...: 5439 4461 8862 2335 8111 1847 716 1030 6621 2368 ...
 $ Codi.districte      : int -1 -1 10 10 10 10 10 10 10 10 ...
 $ Codi.barri          : int -1 -1 64 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 ...
 $ Dia.setmana         : Factor w/ 7 levels "Dc","Dg","Dj",...: 5 3 5 5 1 5 2 7 5 1 ...
 $ Descripció.tipus.dia : Factor w/ 1 level "Laboral": 1 1 1 1 1 1 1 1 1 1 ...
 $ NK.Any              : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ Mes.de.any          : int 8 6 12 4 11 3 2 2 9 4 ...
 $ Nom.mes             : Factor w/ 12 levels "Abril","Agost",...: 2 7 3 1 10 9 4 4 12 1 ...
 $ 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.implicat : Factor w/ 19 levels "Autobús","Autobús articulado",...: 13 19 13 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                : Factor w/ 98 levels "0","1","10","101",...: 33 53 47 8 47 68 30 34 47 19 ...
 $ Descripció.victimització : Factor w/ 3 levels "Ferit greu","Ferit lleu",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ Coordenada.UTM..Y.   : Factor w/ 5612 levels "-1","4575062,68",...: 1 1 4249 4404 4287 4267 4275 4316 4279 4386 ...
 $ Coordenada.UTM..X.   : Factor w/ 5612 levels "-1","423577,48",...: 1 1 4601 4445 4258 4321 4319 4306 4200 4266 ...
```

## 3 – Limpieza de datos

### 3.1 Búsqueda valores vacíos o nulos

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

```
summarise_all(anyo2015, funs(sum(is.na(.))))
```

```
• summarise_all(anyo2015, funs(sum(is.na(.))))
Número.d.expedient Codi.districte Codi.barri Codi.carrer Num.postal.caption Descripció.dia.setmana Dia.setmana Descripció.tipus.dia NK.Any Mes.de.any
0 0 0 0 0 0 0 0 0 0 0 0
Nom.mes Dia.de.mes Hora.de.dia Descripció.causa.vianant Desc.Tipus.vehicle.implicat Descripció.sexe Descripció.tipus.persona Edat Descripció.victimització
0 0 0 0 0 0 0 0 0 0
Coordenada.UTM..Y. Coordenada.UTM..X.
0 0
```

NO existen valores NA pero si hemos encontrado campos con el valor “-1”  
(Desconocido)

Vamos a considerar los accidentes con valores desconocidos como no validos y los vamos a eliminar del dataset.

### 3.2 Identificación y tratamiento de valores extremos.

Vamos a analizar las variables a ver si existen valores extremos o erróneos, descartando las variables que contienen los códigos de barrios, distritos y calles de Barcelona que se entienden que son correctas.

La mayoría de las variables son categoricas, así que vamos a comprobar que sus valores sean útiles, no erróneos o nulos.

#### - Descripcio.dia.setmana

La columna contiene valores correctos con los días de la semana.

```
> table(anyo2015$Descripció.dia.setmana)
Dijous    Dilluns    Dimarts    Dimecres    Dissabte    Diumenge    Divendres
1961      1662      1826      1934      1308      1033      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 **descarta**.

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

```
Laboral  
11780
```

#### - Nom.mes

Los valores de esta variable son correctos, entre el rango [1-12]

```
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
```

#### - Dia.de.mes

Los valores de esta variable son correctos, se encuentran en el rango [1-31]

```
> #Valores día 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

Otra variable categórica, los datos son correctos.

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

	Altres Creuar per fora pas de vianants	Desobeir altres senyals	Desobeir el senyal del semàfor
	123	271	1
No és causa del vianant	Transitar a peu per la calçada		317
	11028	40	

```
> |
```

## - Descripcion.tipus.vehiculo.implicat

Variable categórica con los tipos de vehículos., sin valores anormales.

```
> #Descripcion tipo vehiculos
```

```
> table(anyo2015$Desc..Tipus.vehicle.implicat)
```

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

```
> |
```

## - Descripcion.sexe

Variable sobre el sexo de las personas siniestradas.

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

```
> #Descripcion sexo
```

```
> table(anyo2015$Descripció.sexe)
```

Desconegut	Dona	Home
3	4405	7372

```
> |
```

## - Descripcio.tipus.persona

Variable categórica, los resultados son correctos.

```
> #Descripcion 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” 124 veces.

```
> table(ano2015$Edat)
```

```

0      1      10     101     106     11     12     13     14     15     16     17     18     19
24     13     26      1      1     29     30     30     41     39     43     98    158    168
2      20     21     22     23     24     25     26     27     28     29      3     30     31
16    229    237    283    306    324    350    340    311    344    286    25    323    338
32     33     34     35     36     37     38     39      4     40     41     42     43     44
325    298    320    292    291    277    282    296    25    246    261    247    205    238
45     46     47     48     49      5     50     51     52     53     54     55     56     57
231    191    170    189    184     32    184    176    170    139    148    115    141    114
58     59      6     60     61     62     63     64     65     66     67     68     69      7
92    103     23     99     78     75     69     58     44     41     47     53     45     21
70     71     72     73     74     75     76     77     78     79      8     80     81     82
55     40     33     34     25     32     16     31     24     34     27     27     37     25
83     84     85     86     87     88     89      9     90     91     92     94     96 Desconegut
27     20     10     15     20     15      8     22     10      9      6      4      1     125

```

Una vez eliminados estos valores obtenemos este histograma, donde apreciamos el valor 0 como mínimo y el 106 como máximo.

## - Descripció.victimització

Variable categórica con valores correctos.

```
> #Descripcion victimas
> table(ano2015$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.

## 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(Anyo2015$Descripció.victimització,Anyo2015$Nom.mes)
```

```
      Abril Agost Desembre Febrer Gener Juliol Juny Maig Març Novembre Octubre Setembre
Ferit greu    15     17      20     15     12      13     18     23     12      23     17      12
Ferit lleu   992    774   1033    876    921    991 1002 1000   993   1058   1013    904
Mort          1      0       3       4       1       2      3      3      1       3       1       4
```

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

```
> prop.table(table(Anyo2015$Descripció.victimització,Anyo2015$Nom.mes),1)
```

```
      Abril Agost Desembre Febrer Gener Juliol Juny Maig Març Novembre Octubre
Ferit greu 0.07614213 0.08629442 0.10152284 0.07614213 0.06091371 0.06598985 0.09137056 0.11675127 0.06091371 0.11675127 0.08629442
Ferit lleu 0.08583542 0.06697240 0.08938306 0.07579822 0.07969196 0.08574890 0.08670070 0.08652765 0.08592195 0.09154625 0.08765250
Mort       0.03846154 0.00000000 0.11538462 0.15384615 0.03846154 0.07692308 0.11538462 0.11538462 0.03846154 0.11538462 0.03846154

      Setembre
Ferit greu 0.06091371
Ferit lleu 0.07822099
Mort       0.15384615
```



Heridos por meses

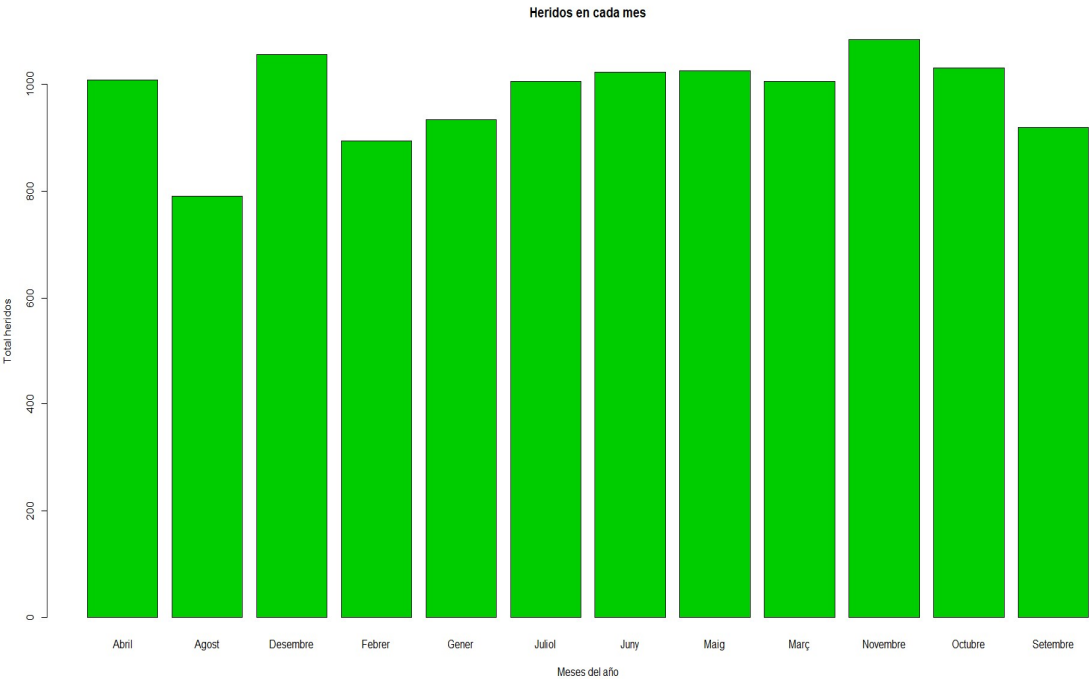


Tabla de frecuencia cruzada días del mes y gravedad heridas

```
> table(anyo2015$Descripció.victimització,anyo2015$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
Ferit greu	4	5	4	7	9	3	8	3	13	4	8	5	5	10	10	5	8	6	11	4	11	5	6	7	4	6	5	5	7	4
Ferit lleu	278	387	396	407	372	374	333	395	437	402	353	328	403	350	394	431	413	358	419	392	374	400	410	366	369	351	383	393	329	367
Mort	0	1	0	0	0	1	2	0	1	1	2	1	1	1	1	0	0	1	3	1	1	1	2	0	2	0	1	1	0	1

	31
Ferit greu	5
Ferit lleu	193
Mort	0

Probabilidad tabla de frecuencia cruzada días del mes y gravedad heridas

```
> prop.table(table(anyo2015$Descripció.victimització,anyo2015$Dia.de.mes),1)
```

	1	2	3	4	5	6	7	8	9	10	11
Ferit greu	0.02030457	0.02538071	0.02030457	0.03553299	0.04568528	0.01522843	0.04060914	0.01522843	0.06598985	0.02030457	0.04060914
Ferit lleu	0.02405469	0.03348620	0.03426495	0.03521675	0.03218828	0.03236134	0.02881371	0.03417842	0.03781258	0.03478411	0.03054426
Mort	0.00000000	0.03846154	0.00000000	0.00000000	0.00000000	0.03846154	0.07692308	0.00000000	0.03846154	0.03846154	0.07692308

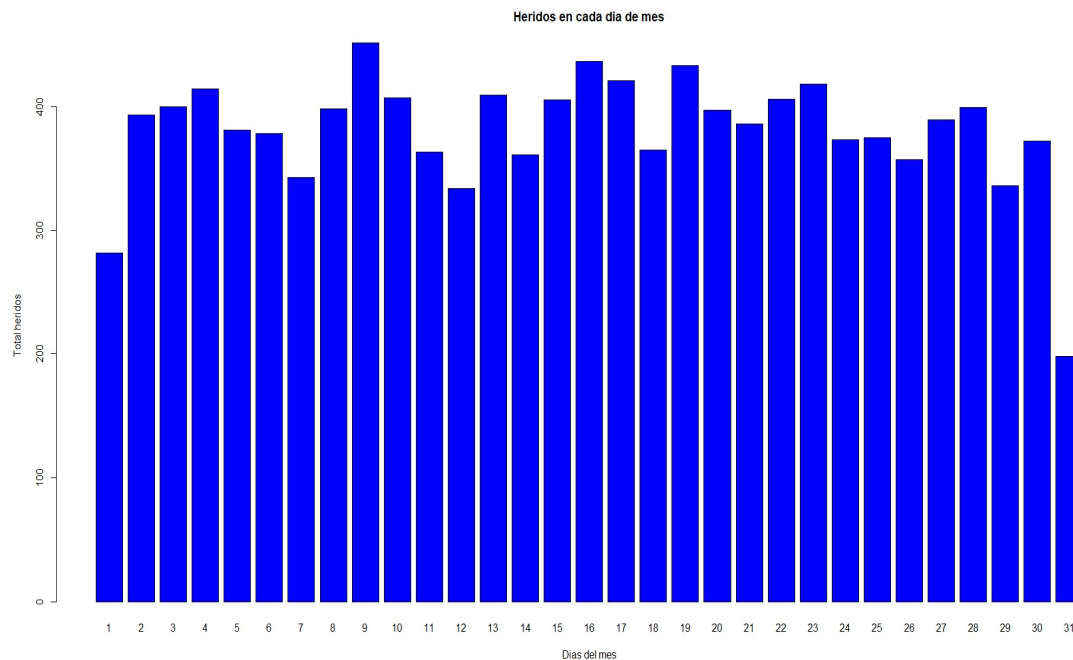
  

	12	13	14	15	16	17	18	19	20	21	22
Ferit greu	0.02538071	0.02538071	0.05076142	0.05076142	0.02538071	0.04060914	0.03045685	0.05583756	0.02030457	0.05583756	0.02538071
Ferit lleu	0.02838107	0.03487064	0.03028468	0.03409189	0.03729342	0.03573592	0.03097690	0.03625508	0.03391884	0.03236134	0.03461106
Mort	0.03846154	0.03846154	0.03846154	0.03846154	0.00000000	0.00000000	0.03846154	0.11538462	0.03846154	0.03846154	0.03846154

	23	24	25	26	27	28	29	30	31
Ferit greu	0.03045685	0.03553299	0.02030457	0.03045685	0.02538071	0.02538071	0.03553299	0.02030457	0.02538071
Ferit lleu	0.03547633	0.03166912	0.03192870	0.03037120	0.03314009	0.03400536	0.02846760	0.03175565	0.01669984
Mort	0.07692308	0.00000000	0.07692308	0.00000000	0.03846154	0.03846154	0.00000000	0.03846154	0.00000000

## Heridos por día del mes



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

```
> table(anyo2015$Descripció.victimització,anyo2015$Dia.setmana)
```

	Dc	Dg	Dj	Di	Dm	Ds	Dv
Ferit greu	35	15	27	24	31	32	33
Ferit lleu	1894	1016	1932	1635	1791	1272	2017
Mort	5	2	2	3	4	4	6

```
> |
```

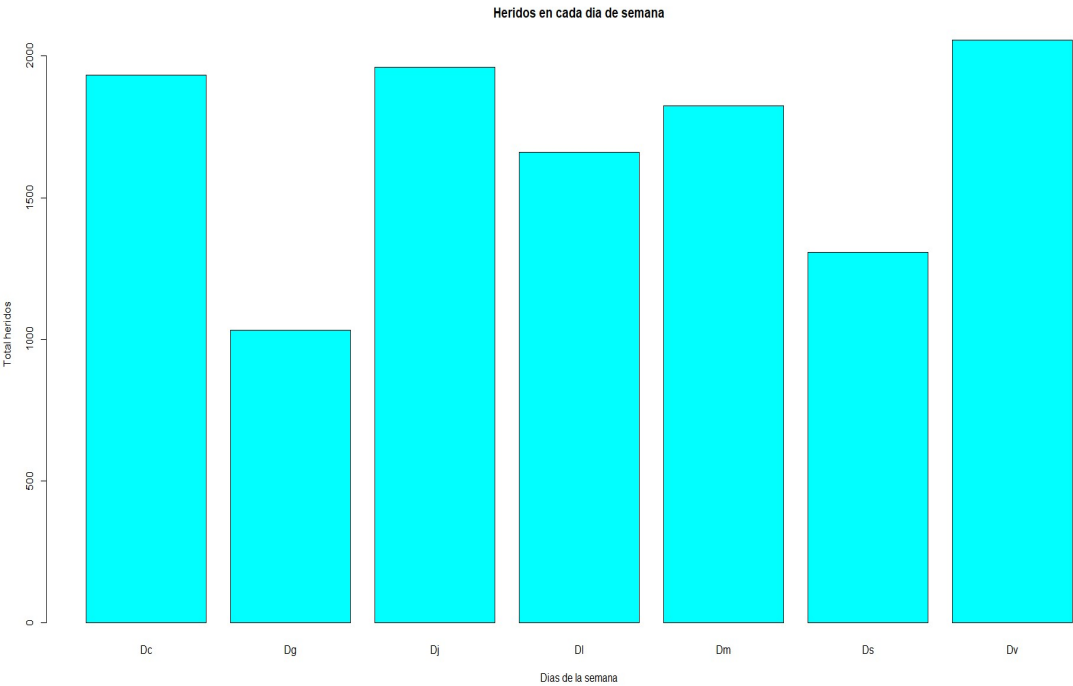
## Probabilidad tabla de frecuencia cruzada días de la semana y gravedad heridas

```
> prop.table(table(anyo2015$Descripció.victimització,anyo2015$Dia.setmana),1)
```

	Dc	Dg	Dj	Di	Dm	Ds	Dv
Ferit greu	0.17766497	0.07614213	0.13705584	0.12182741	0.15736041	0.16243655	0.16751269
Ferit lleu	0.16388336	0.08791209	0.16717141	0.14147270	0.15497101	0.11006317	0.17452626
Mort	0.19230769	0.07692308	0.07692308	0.11538462	0.15384615	0.15384615	0.23076923

```
> |
```

# Heridos por día de la semana



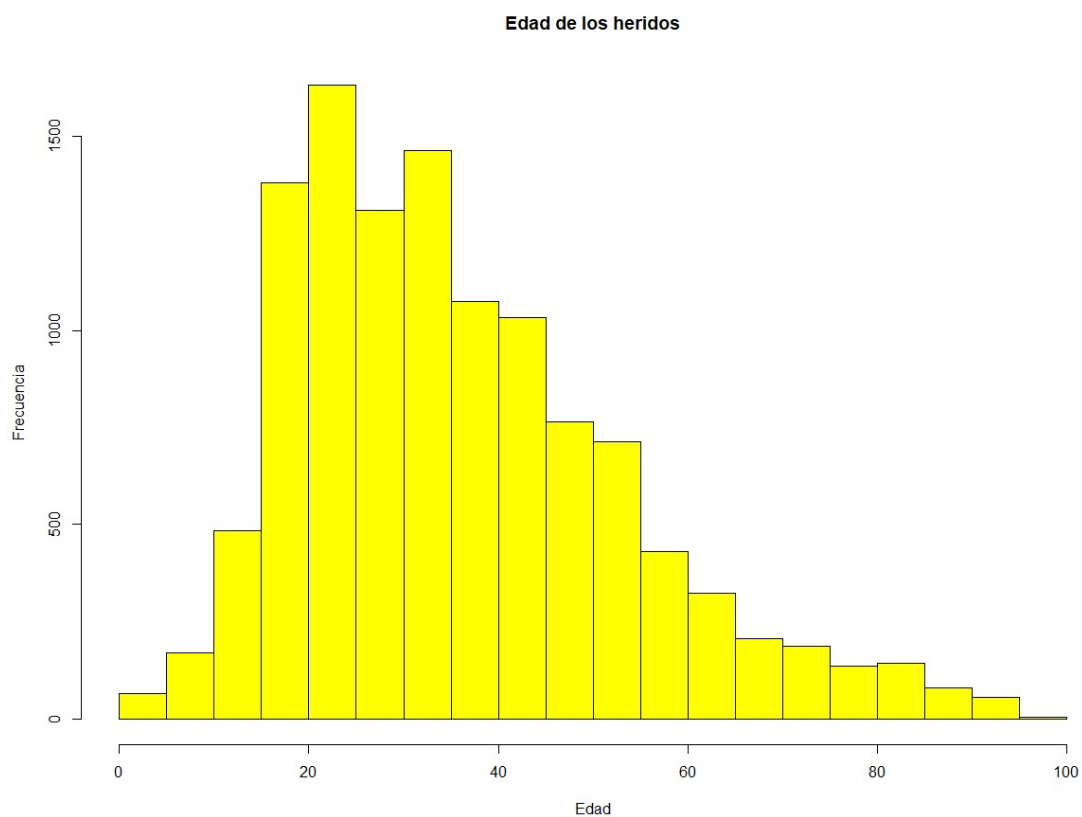
## 4.2 Análisis social de los heridos.

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 98.

A continuación, el Histograma de las edades.

### Edad de los heridos

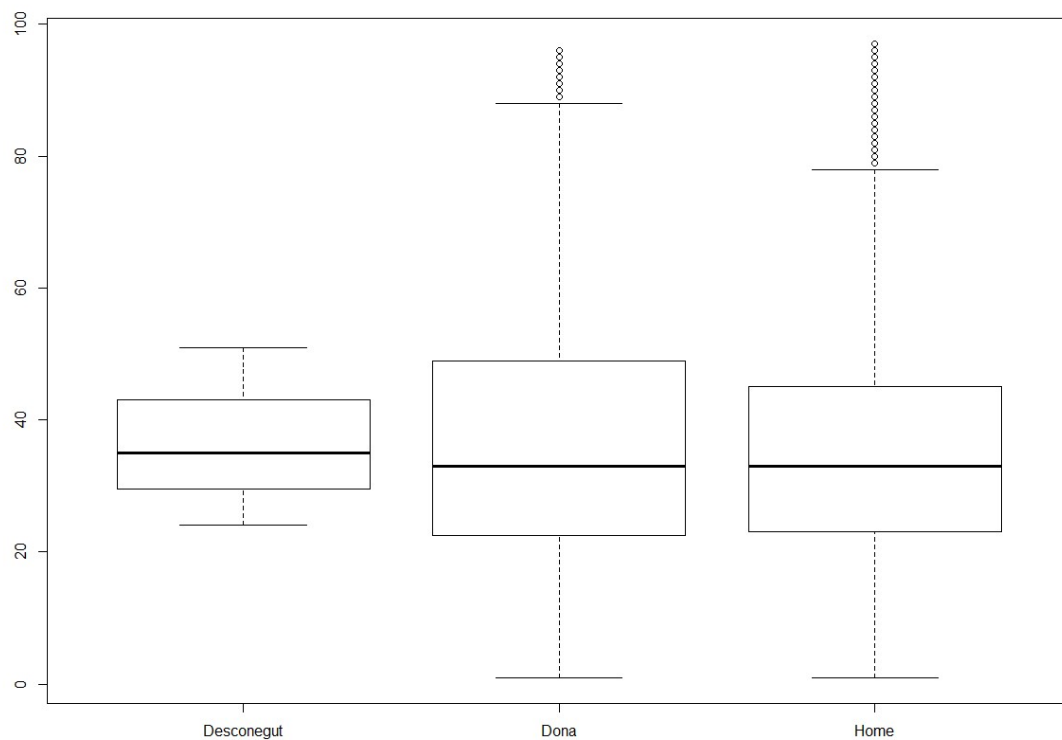


### Frecuencias por edad y sexo

```
> #Frecuencias edad y sexo
> edadsexo= data.frame(anyo2015Edat$Edat, anyo2015Edat$Descripció.sexe)
> summary(edadsexo)
anyo2015Edat.Edat anyo2015Edat.Descripció.sexe
Min. : 1.00      Desconegut: 3
1st Qu.: 23.00    Dona : 4364
Median : 33.00    Home : 7288
Mean : 36.33
3rd Qu.: 46.00
Max. : 97.00
> |
```

## Caculo de la desviación edad por sexo

```
> aggregate(anyo2015Edat$Edat,by=list(anyo2015Edat$Descripció.sexe),mean,na.rm=TRUE)
  Group.1      x
1 Desconegut 36.66667
2      Dona  37.36732
3      Home  35.71144
> |
```



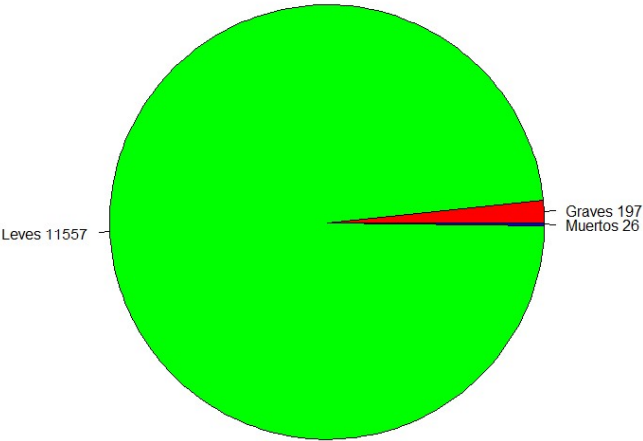
## Relación gravedad heridas y sexo

```
> #Relacion heridas y sexo
> table(anyo2015$Descripció.victimització,anyo2015$Descripció.sexe)

      Desconegut Dona Home
Ferit greu      0   65  132
Ferit lleu      3 4338 7216
Mort            0    2   24
> |
```

Tipos y cantidad de heridos

Tipos y cantidad de heridos



### 4.3 Análisis geoespacial de los accidentes.

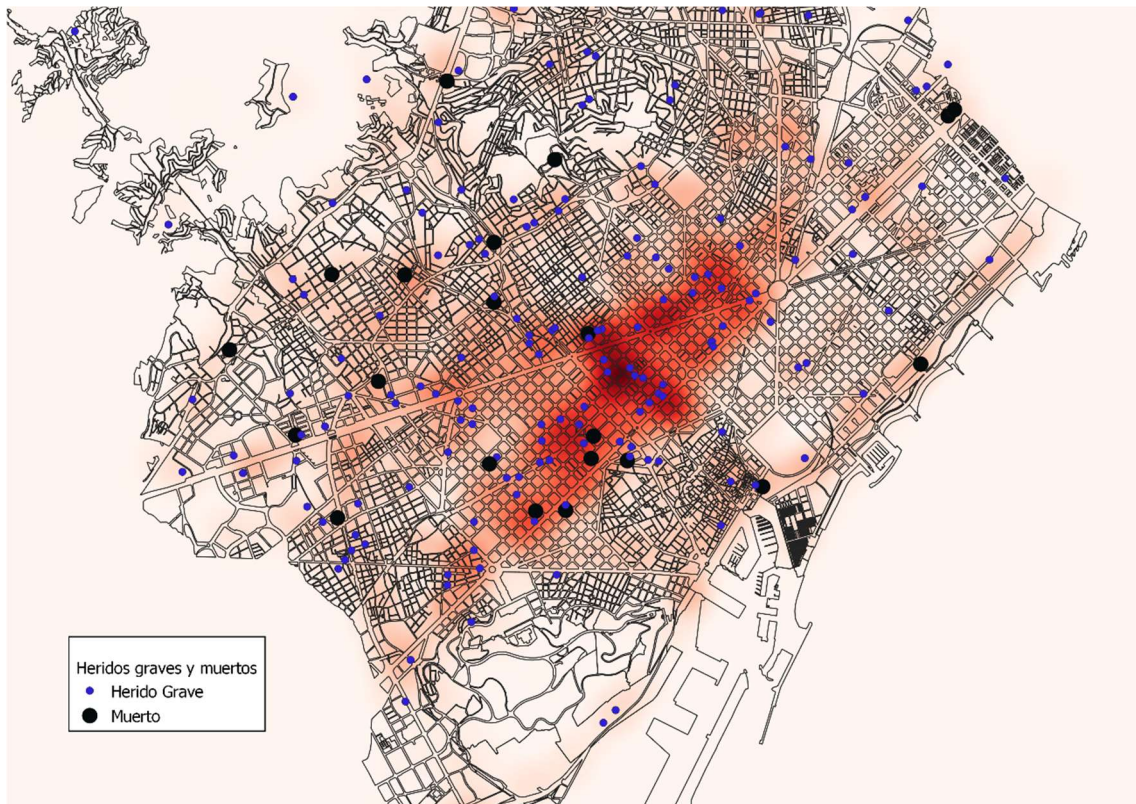
Antes de empezar el análisis vamos a eliminar los heridos no georreferenciados, con valores de '-1'.

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.

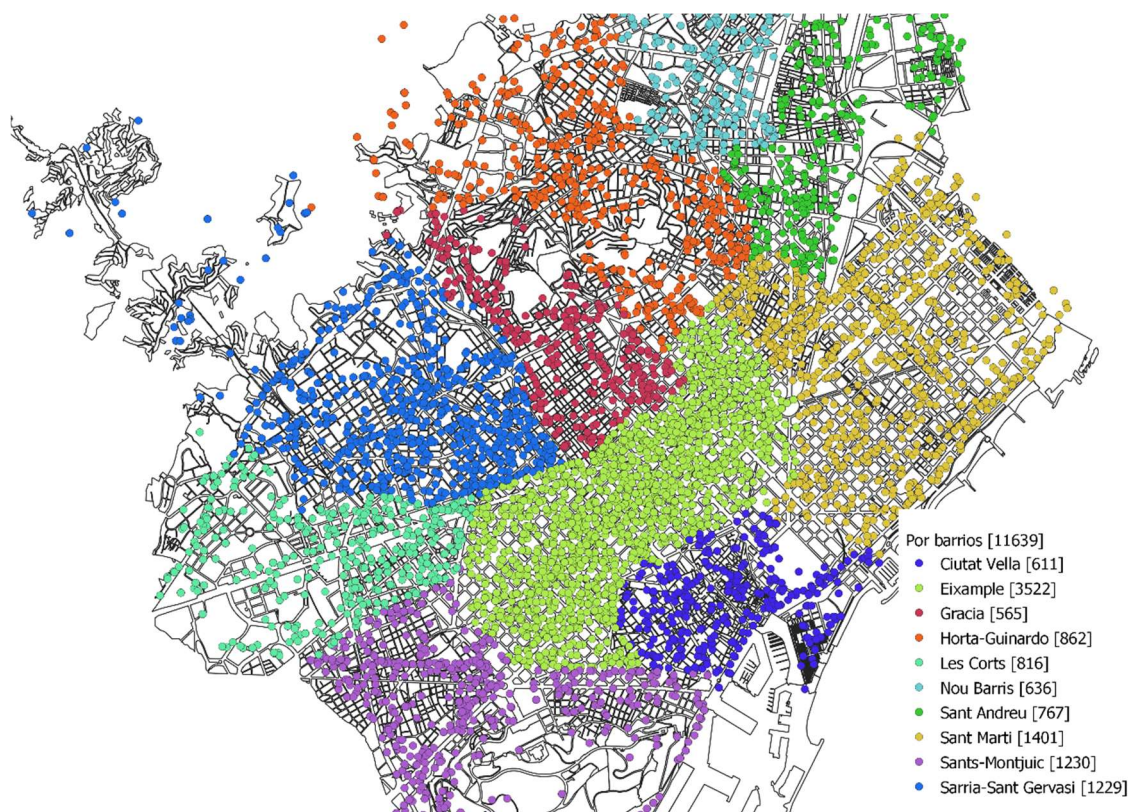
#### Mapa de calor de los accidentes sobre cartografía Barcelona

Mapa de calor con posicionamiento de los heridos graves y muertos

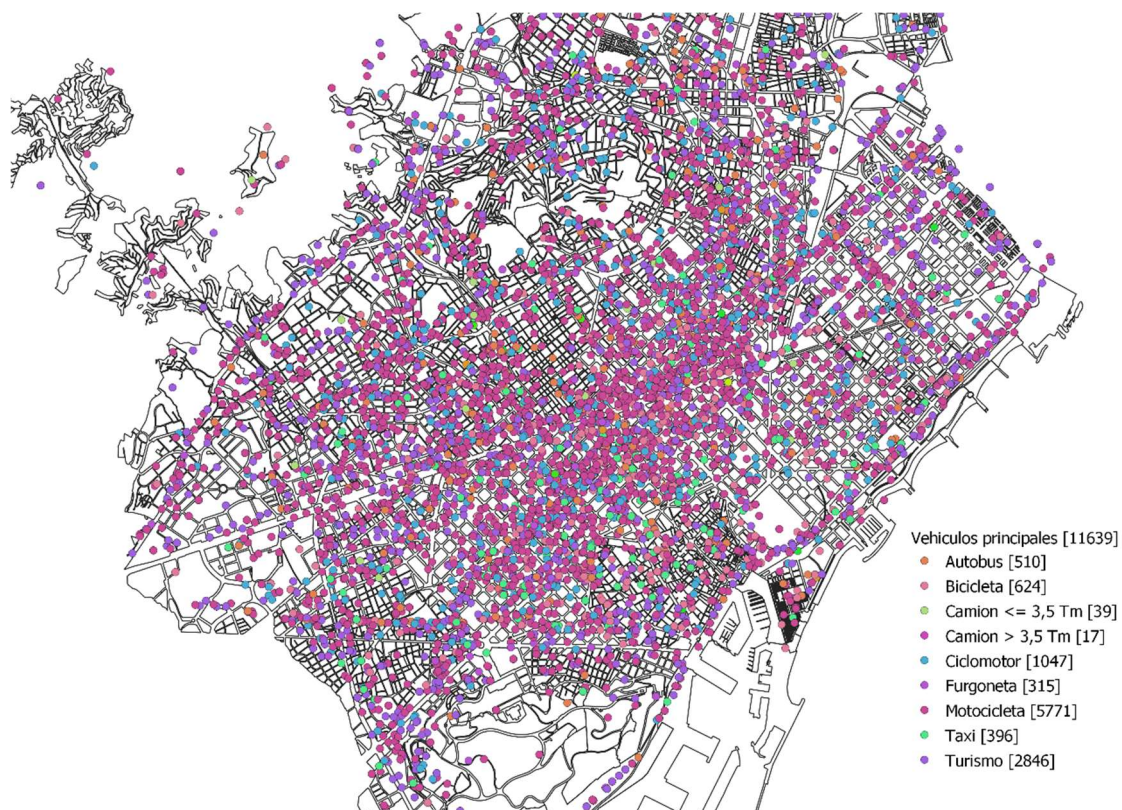




## Mapa de los accidentes por barrios



## 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.

Los meses de mas actividad son los de noviembre y diciembre y en agosto, suponemos que por las vacaciones y el descenso del tráfico, es el mes con menos heridos.

A lo largo del mes los accidentes están repartido de forma homogénea por quincenas y el día de la semana mas peligroso para circular por Barcelona es el viernes.

En cuanto a los heridos, el perfil del herido es de un hombre de edad entre 20-40 años.

Los accidentes se producen por toda Barcelona, pero la zona más conflictiva, según el mapa de calor, es la zona del Eixample.

## 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/marcvelfer/barcelona-accident> procedente de <http://opendata-ajuntament.barcelona.cat/en/>