Práctica 2: Estudio de los accidentes de tráfico de Barcelona durante el año 2018.

Marcos Pereiro Conde (Universitat Oberta de Catalunya)

30 de diciembre de 2019

Contents

1 Introducción

En este proyecto se analizan los accidentes de tráfico ocurridos en la ciudad de Barcelona durante el año 2018. El Ayuntamiento de Barcelona, a través de su portal de Open Data, publica anualmente los datos de los accidentes gestionados por la Guardia Urbana durante cada periodo.

Se usarán esos datos para caracterizar los accidentes de tráfico en la ciudad, poniendo especial énfasis en determinar qué factores influyen en mayor medida en la gravedad del accidente. Conocer esos factores puede ayudar a las administraciones y organismos a establecer medidas que prevengan o palien las consecuencias de un accidente.

2 Descripción de los datasets

La información se puede obtener a través de varios ficheros que representan distintas dimensiones de los accidentes. Por un lado encontramos el fichero '2018_accidents_gu_bcn.csv' que contiene los datos de los accidentes en sí mismos. En el fichero '2018_accidents_persones_gu_bcn_.csv' encontramos los datos de las personas accidentadas y en el fichero '2018_accidents_tipus_gu_bcn_.csv' encontramos la tipología de cada accidente (colisión frontal, alcance, salida de vía, etc.)

Fichero "Accidentes" (2018_accidents_gu_bcn.csv)

Este dataset de accidentes encontramos:

- **Numero_expedient**: Identificador del expediente del accidente
- **Codi_districte**: Código numérico del distrito
- **Nom_districte**: Nombre del distrito
- **Codi_barri**: Código numérico del barrio
- **Nom_barri**: Nombre del barrio
- **Codi_carrer**: Código numérico de la calle
- **Nom_carrer**: Nombre de la calle
- **Num_postal**: Número de la calle
- **Descripcio_dia_setmana**: Descripción del día de la semana
- **Dia_setmana**: Código del día de la semana
- **Descripcio_tipus_dia**: Tipo de día (laboral)
- **Any**: Año del accidente
- **Mes_any**: Mes del accidente, en número
- **Nom_mes**: Nombre del mes del accidente
- **Dia_mes**: Día del mes del accidente
- **Hora_dia**: Hora del accidente
- **Descripcio_torn**: Descripción del turno (matí, tarda, nit)
- **Descripcio_causa_vianant**: Descripción de la causa del accidente, si es por motivo de un viandante
- **Numero_morts**: Número de muertos en el accidente
- **Numero_lesionats_lleus**: Número de lesionados leves
- **Numero_lesionats_greus**: Número de lesionados graves
- **Numero_victimes**: Número de víctimas (leves y graves)
- **Numero_vehicles_implicats**: Número de vehículos implicados
- **Coordenada_UTM_X**: Coordenada X de la localización del accidente en estandar UTM
- **Coordenada_UTM_Y**: Coordenada Y de la localización del accidente en estandar UTM
- **Longitud**: Coordenada del accidente (longitud)
- **Latitud**: Coordenada del accidente (latitud)

Fichero "Personas" (2018_accidents_persones_gu_bcn_.csv)

Contiene en parte los mismos campos que el fichero de accidentes (los que hacen referencia al expediente, datos de localización y fecha). Los campos que difieren son:

- **Desc_Tipus_vehicle_implicat**: Descripción del tipo de vehiculo accidentado (turismo, ciclomotor, ...)
- **Descripcio_sexe**: Descripción del sexo de la persona accidentada
- **Edat**: Edad de la persona accidentada
- **Descripcio_tipus_persona**: Tipo de persona (conductor, pasajero, viandante)
- **Descripcio_victimitzacio**: Gravedad del accidentado (leve, grave, muerto)

Fichero "Tipología de accidente" (2018_accidents_tipus_gu_bcn_.csv)

Este fichero también contiene en parte los mismos campos que el fichero de accidentes (los que hacen referencia al expediente, datos de localización y fecha). Se añade un campo:

- **Tipus_accident**: Descripción del tipo de accidente (colisión frontal, colisión lateral, alcance por detras, etc.)

La clave que permite enlazar estos datasets es el número de expediente del accidente. Los registros pueden encontrarse duplicados: por ejemplo, un accidente con dos tipologías de accidente (colisión frontal y lateral) aparecerá dos veces en el fichero de "Tipología de accidentes".

3 Carga de datos e instalación de librerías

```
# Instala librerías si no están ya instaladas
if(!require("knitr")) install.packages("knitr")
if(!require("dplyr")) install.packages("dplyr")
if(!require("ggmap")) install.packages("ggmap")
if(!require("osmdata")) install.packages("osmdata")
if(!require("tidyverse")) install.packages("tidyverse")
if(!require("lsr")) install.packages("lsr")
if(!require("ROCR")) install.packages("ROCR")

# Carga librerías
library(knitr)
library(dplyr)
library(ggmap)
library(osmdata)
library(tidyverse)
library(lsr)
library(ROCR)

# Lectura de datos

# Datos de accidentes
accidentes <-read.csv("2018_accidents_gu_bcn.csv",head=TRUE,sep=",", encoding="UTF-8",
                 stringsAsFactors = FALSE)

# Tipo de accidente
tipos <-read.csv("2018_accidents_tipus_gu_bcn_.csv",head=TRUE,sep=",", encoding="UTF-8",
                 stringsAsFactors = FALSE)
```

```r
# Datos de víctimas
pers <-read.csv("2018_accidents_persones_gu_bcn_.csv",head=TRUE,sep=",", encoding="UTF-8",
                stringsAsFactors = FALSE)


# Listamos una muestra de registros de cada dataset y su estructura
head(accidentes, 3)
```

```
##   Numero_expedient Codi_districte  Nom_districte Codi_barri
## 1  2018S000150                 3 Sants-Montjuïc         12
## 2  2018S000761                 3 Sants-Montjuïc         11
## 3  2018S005151                 3 Sants-Montjuïc         12
##                    Nom_barri Codi_carrer
## 1 la Marina del Prat Vermell      701819
## 2               el Poble-sec      234001
## 3 la Marina del Prat Vermell      370531
##                                        Nom_carrer Num_postal
## 1 Número 3 Zona Franca / E Zona Franca                69-75
## 2 Olímpic                                             5-7
## 3 A Zona Franca / Número 6 Zona Franca
##   Descripcio_dia_setmana Dia_setmana Descripcio_tipus_dia  Any Mes_any Nom_mes
## 1                Dilluns          Dl              Laboral 2018       1   Gener
## 2               Diumenge          Dg              Laboral 2018       1   Gener
## 3               Dimecres          Dc              Laboral 2018       7  Juliol
##   Dia_mes Hora_dia Descripcio_torn Descripcio_causa_vianant Numero_morts
## 1       8        7            Matí No és causa del  vianant            0
## 2      28       19           Tarda No és causa del  vianant            0
## 3       4       13            Matí No és causa del  vianant            0
##   Numero_lesionats_lleus Numero_lesionats_greus Numero_victimes
## 1                      1                      0               1
## 2                      0                      0               0
## 3                      1                      0               1
##   Numero_vehicles_implicats Coordenada_UTM_X Coordenada_UTM_Y Longitud  Latitud
## 1                         2         427266.0          4576645 2.129612 41.33616
## 2                         2         429210.6          4579530 2.152513 41.36233
## 3                         2         427520.0          4575229 2.132817 41.32344
```

```r
str(accidentes)
```

```
## 'data.frame': 9936 obs. of 27 variables:
## $ Numero_expedient : chr "2018S000150 " "2018S000761 " "2018S005151 " "2018S000933 "
...
## $ Codi_districte : int 3 3 3 3 3 1 3 3 3 3 ...
## $ Nom_districte : chr "Sants-Montjuïc" "Sants-Montjuïc" "Sants-Montjuïc"
"Sants-Montjuïc" ...
## $ Codi_barri : int 12 11 12 12 12 4 12 12 12 12 ...
## $ Nom_barri : chr "la Marina del Prat Vermell" "el Poble-sec" "la Marina del Prat
Vermell" "la Marina del Prat Vermell" ...
## $ Codi_carrer : int 701819 234001 370531 701819 370203 207803 370531 701305 370531
370407 ...
## $ Nom_carrer : chr "Número 3 Zona Franca / E Zona Franca " "Olímpic " "A Zona Franca /
Número 6 Zona Franca " "Número 3 Zona Franca / E " ...
## $ Num_postal : chr "69-75 " "5-7 " "" "69-75 " ...
```

```
## $ Descripcio_dia_setmana : chr "Dilluns" "Diumenge" "Dimecres" "Diumenge" ...
## $ Dia_setmana : chr "Dl" "Dg" "Dc" "Dg" ...
## $ Descripcio_tipus_dia : chr "Laboral" "Laboral" "Laboral" "Laboral" ...
## $ Any : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
## $ Mes_any : int 1 1 7 2 9 3 9 10 12 9 ...
## $ Nom_mes : chr "Gener" "Gener" "Juliol" "Febrer" ...
## $ Dia_mes : int 8 28 4 4 14 19 21 26 13 20 ...
## $ Hora_dia : int 7 19 13 7 6 9 6 18 6 23 ...
## $ Descripcio_torn : chr "Matí" "Tarda" "Matí" "Matí" ...
## $ Descripcio_causa_vianant : chr "No és causa del vianant" "No és causa del vianant"
"No és causa del vianant" "No és causa del vianant" ...
## $ Numero_morts : int 0 0 0 0 0 0 0 0 0 0 ...
## $ Numero_lesionats_lleus : int 1 0 1 2 0 0 1 1 1 1 ...
## $ Numero_lesionats_greus : int 0 0 0 0 0 0 0 0 0 1 ...
## $ Numero_victimes : int 1 0 1 2 0 0 1 1 1 2 ...
## $ Numero_vehicles_implicats: int 2 2 2 1 2 1 2 2 2 2 ...
## $ Coordenada_UTM_X : num 427266 429211 427520 427361 426842 ...
## $ Coordenada_UTM_Y : num 4576645 4579530 4575229 4576569 4576816 ...
## $ Longitud : num 2.13 2.15 2.13 2.13 2.12 ...
## $ Latitud : num 41.3 41.4 41.3 41.3 41.3 ...
```

```
# Variables de interés en dataset de tipologia de accidente
vtip =c("Codi_expedient", "Tipus_accident")

head(tipos[,vtip], 3)
```

```
##    Codi_expedient           Tipus_accident
## 1 2018S000001               Abast multiple
## 2 2018S000002     Col.lisió fronto-lateral
## 3 2018S000003                         Abast
```

```
str(tipos[,vtip])
```

```
## 'data.frame': 10551 obs. of 2 variables:
## $ Codi_expedient: chr "2018S000001 " "2018S000002 " "2018S000003 " "2018S000004 " ...
## $ Tipus_accident: chr "Abast multiple" "Col.lisió fronto-lateral" "Abast" "Xoc contra
element estàtic" ...
```

```
# Variables de interés en dataset de personas
vpers =c("Numero_expedient", "Desc_Tipus_vehicle_implicat", "Descripcio_sexe",
              "Edat", "Descripcio_tipus_persona", "Descripcio_victimitzacio")

head(pers[,vpers], 3)
```

```
##   Numero_expedient Desc_Tipus_vehicle_implicat Descripcio_sexe Edat
## 1  2018S000001                       Turisme            Home   41
## 2  2018S000001                       Turisme            Home   54
## 3  2018S000002                    Ciclomotor            Home   45
##   Descripcio_tipus_persona
## 1               Conductor
## 2               Conductor
## 3               Conductor
```

```
##                                 Descripcio_victimitzacio
## 1 Ferit lleu: Amb assistència sanitària en lloc d'accident
## 2 Ferit lleu: Amb assistència sanitària en lloc d'accident
## 3                  Ferit lleu: Hospitalització fins a 24h
```

```
str(pers[,vpers])
```

```
## 'data.frame': 11854 obs. of 6 variables:
## $ Numero_expedient : chr "2018S000001 " "2018S000001 " "2018S000002 " "2018S000003 "
...
## $ Desc_Tipus_vehicle_implicat: chr "Turisme" "Turisme" "Ciclomotor" "Motocicleta" ...
## $ Descripcio_sexe : chr "Home" "Home" "Home" "Home" ...
## $ Edat : chr "41" "54" "45" "70" ...
## $ Descripcio_tipus_persona : chr "Conductor" "Conductor" "Conductor" "Conductor" ...
## $ Descripcio_victimitzacio : chr "Ferit lleu: Amb assistència sanitària en lloc
d'accident" "Ferit lleu: Amb assistència sanitària en lloc d'accident" "Ferit lleu:
Hospitalització fins a 24h" "Ferit lleu: Hospitalització fins a 24h" ...
```

4 Preprocesamiento de datos

4.1 Selección de campos

En primer lugar procederemos a eliminar aquellas variables que no son relevantes para nuestro problema. En este estudio no se pretende analizar los accidentes por su localización, por lo que eliminaremos aquellas variables relativas a la ubicación. También eliminaremos aquellas que hacen referencia al tiempo (dia, mes, ...) salvo el día de la semana. La variable "Descripcio_causa_vianant" sólo se informa en un porcentaje muy pequeño de casos, por lo que tampoco la usaremos en nuestro análisis.

```
#
# Dataset Accidentes:
#
# Eliminamos las variables del dataset que no vamos a utilizar en este estudio

accidentes <- accidentes[,-which(names(accidentes) %in% c("Descripcio_causa_vianant",
    "Codi_districte", "Codi_barri", "Nom_barri", "Codi_carrer", "Nom_carrer",
    "Num_postal", "Dia_setmana", "Any","Mes_any", "Nom_mes", "Dia_mes", "Hora_dia",
    "Numero_morts", "Numero_lesionats_lleus", "Numero_lesionats_greus", "Numero_victimes",
    "Numero_vehicles_implicats", "Coordenada_UTM_X", "Coordenada_UTM_Y",
    "Descripcio_tipus_dia"))]

# Elimina registros duplicados tras la eliminación de columnas
accidentes <- accidentes %>% distinct()



#
# Dataset Tipos de accidentes:
#
# Seleccionamos las variables de interés del dataset tipos. El resto están en accidentes
tipos <- tipos[c("Codi_expedient", "Tipus_accident")]

# Elimina registros duplicados tras la eliminación de columnas
```

```r
tipos <- tipos %>% distinct()


#
# Dataset Personas (víctimas):
#
# Seleccionamos las variables de interés. El resto están en accidentes
pers <- pers[c("Numero_expedient", "Desc_Tipus_vehicle_implicat", "Descripcio_sexe",
               "Edat", "Descripcio_tipus_persona", "Descripcio_victimitzacio")]

# Elimina registros duplicados tras la eliminación de columnas
pers <- pers %>% distinct()
```

4.2 Unión de datasets

Creamos un nuevo dataset a partir de los datasets de accidentes y personas, unidos a través del código de expediente. Esta es una relación de 1 a N y el fichero resultante contiene, por tanto, las personas accidentadas.

```r
acc <- merge (accidentes, pers)
```

4.3 Abreviar nombres de variables

Cambiamos el nombre de las variables por otros más cortos para facilitar su utilización.

```r
newnames <- c("Exp","Dist","Dia", "Turno", "Long", "Lat", "TipVeh", "Sexo", "Edad",
              "TipPer", "Vict")
colnames(acc) <- newnames

colnames(tipos) <- c("Exp","TipAcc")
```

4.4 Limpieza de valores perdidos o desconocidos

Examinamos primero los valores que toman las distintas variables de tipo carácter, para ver si observamos valores perdidos o desconocidos.

```r
sapply(acc[,c("Dist","Dia","Turno", "TipVeh", "Sexo", "Vict")],
       function(x) kable(sort(table(x), decreasing = TRUE)) )
```

```
## $Dist
##
##
## x                     Freq
## -------------------  -----
## Eixample              3574
## Sant Martí            1438
## Sarrià-Sant Gervasi   1321
## Sants-Montjuïc        1229
## Horta-Guinardó         855
## Les Corts              848
```

```
## Sant Andreu             767
## Nou Barris              638
## Ciutat Vella            618
## Gràcia                  554
##
## $Dia
##
##
## x            Freq
## ----------  -----
## Divendres    2011
## Dimarts      1871
## Dimecres     1849
## Dijous       1840
## Dilluns      1811
## Dissabte     1400
## Diumenge     1060
##
## $Turno
##
##
## x        Freq
## ------  -----
## Tarda    5925
## Matí     4521
## Nit      1396
##
## $TipVeh
##
##
## x                                    Freq
## ----------------------------------  -----
## Motocicleta                          6018
## Turisme                              2613
## Ciclomotor                            870
## Bicicleta                             710
## Autobús                               553
## Taxi                                  411
## Furgoneta                             333
## Veh. mobilitat personal amb motor     130
## Autobús articulat                      47
## Camió rígid <= 3,5 tones               43
## Camió rígid > 3,5 tones                20
## Altres vehicles amb motor              16
## Tot terreny                            14
## Tren o tramvia                         12
## Veh. mobilitat personal sense motor    12
## Desconegut                             11
## Altres vehicles sense motor             9
## Autocar                                 9
## Maquinària d'obres i serveis            6
## Tractor camió                           3
## Microbús <= 17                          2
##
```

```
## $Sexo
##
##
## x             Freq
## -----------  -----
## Home          7214
## Dona          4624
## Desconegut       4
##
## $Vict
##
##
## x                                                        Freq
## ------------------------------------------------------  -----
## Ferit lleu: Hospitalització fins a 24h                   7577
## Ferit lleu: Amb assistència sanitària en lloc d'accident 3476
## Ferit lleu: Rebutja assistència sanitària                 532
## Ferit greu: hospitalització superior a 24h                236
## Mort (dins 24h posteriors accident)                        21
```

La mayoría de variables están completamente informadas y no contienen valores perdidos. En el caso de la variable Sexo, sí que vemos que hay 4 registros en los que es desconocido. Como necesitaremos este dato para el análisis y son muy pocos registros, procedemos a eliminarlos de la muestra.

```
# Eliminamos aquellos registros que no tienen sexo
acc <- acc[acc$Sexo!="Desconegut",]
```

A continuación, convertimos la variable Edad a numérica. Si hubiera valores no válidos (p.ej. un texto), quedarán como NA. Para no perder estos registros, en estos casos **completaremos el dato mediante la imputación** de la media de edad de la muestra.

```
# Convertimos a numerico. Los valores missing quedan como NA
acc$Edad <- as.numeric(as.character(acc$Edad))

# Imputamos la media de edad a los valores perdidos
media <- floor(mean(acc$Edad, na.rm=TRUE))
acc[is.na(acc$Edad),]$Edad <- media
```

4.5 Outliers

En este dataset sólo hay una variable numérica, edad. Vamos a examinar los outliers mediante la función boxplot.

```
boxplot.stats(acc$Edad)$out
```

```
##   [1] 86 86 85 91 84 86 83 85 86 90 84 92 87 91 89 85 84 97 84 95 84 89 85 98 85
##  [26] 86 90 84 86 83 85 89 89 84 83 88 87 86 86 86 93 88 87 84 89 88 86 83 85 93
##  [51] 91 86 88 87 87 84 85 91 83 91 83 89 85 84 89 86 89 92 83 83 86 93 90 87 87
##  [76] 90 87 84 90 89 83 86 88 85 87 89 90 88 93 84 87 84 84 83 83 87 91 92 86 89
## [101] 95 86 89 83 86 87 86 84 84 87 83 90 88 83 89 88 93 85 88 87 83 91 83 87 83
## [126] 89 88 85 89 90 90 83 91 85 85 87 83 86 92 85 83 83 87 86 84 85 94 84 91 85
## [151] 83 88 85 90 90 87 88 89 94 83 84 86 86 85 83 83 88 85 84 88 84 84 88 83 86
## [176] 83 83 96 89 83 83 90 88 83 91 84
```

Vemos que esos datos son perfectamente posibles: corresponden a personas de edad avanzada. Por tanto, mantendremos estos valores tal cual.

4.6 Creación categoría "Gravedad"

Para simplificar y clarificar el estudio, se creará una nueva categoría "Gravedad" que agrupa los niveles de la variable "Victimizació" y que sólo tomará dos valores: 0 si el accidentado resulta leve y 1 si resulta muerto o grave.

```
acc$gravedad <- ifelse(grepl("greu|Mort", acc$Vict),1,0)
```

4.7 Agrupación categorías "Tipo accidente"

Se reducen los tipos de accidente del dataset original, agrupándolo en unas pocas categorías que reunen accidentes similares.

```
sort(table(tipos$TipAcc), decreasing = TRUE)
```

```
##
##                  Col.lisió lateral                              Abast
##                               2557                               2147
##           Col.lisió fronto-lateral                       Atropellament
##                               1835                               1143
##               Caiguda (dues rodes)          Xoc contra element estàtic
##                                956                                725
##                      Abast multiple            Caiguda interior vehicle
##                                340                                336
##                             Altres                    Col.lisió frontal
##                                256                                158
##        Bolcada (més de dues rodes)                              Encalç
##                                 32                                 30
## Sortida de via amb xoc o col.lisió         Xoc amb animal a la calçada
##                                 17                                 10
##                           Desconegut               Resta sortides de via
##                                  7                                  1
##         Sortida de via amb bolcada
##                                  1
```

```
tipos[ tipos$TipAcc=="Abast multiple" |
    tipos$TipAcc=="Encalç",]$TipAcc <- "Abast"

tipos[ tipos$TipAcc=="Xoc contra element estàtic" |
    tipos$TipAcc=="Xoc amb animal a la calçada" |
    tipos$TipAcc=="Sortida de via amb xoc o col.lisió",]$TipAcc <- "Xoc"

tipos[ tipos$TipAcc=="Bolcada (més de dues rodes)" |
    tipos$TipAcc=="Desconegut" | tipos$TipAcc=="Resta sortides de via"|
    tipos$TipAcc=="Sortida de via amb bolcada",]$TipAcc <- "Altres"


# Elimina filas duplicadas
tipos <- tipos %>% distinct()
```

4.8 Agrupación categorías "Tipo vehiculo"

Se procede de la misma forma que en el punto anterior, agrupando en menos categorías los tipos de vehículos similares.

```
sort(table(acc$TipVeh), decreasing = TRUE)
```

```
##
##                          Motocicleta                              Turisme
##                                 6018                                 2611
##                           Ciclomotor                            Bicicleta
##                                  869                                  709
##                              Autobús                                 Taxi
##                                  553                                  411
##                            Furgoneta   Veh. mobilitat personal amb motor
##                                  333                                  130
##                     Autobús articulat            Camió rígid <= 3,5 tones
##                                   47                                   43
##              Camió rígid > 3,5 tones           Altres vehicles amb motor
##                                   20                                   16
##                           Tot terreny                       Tren o tramvia
##                                   14                                   12
## Veh. mobilitat personal sense motor                          Desconegut
##                                   12                                   11
##          Altres vehicles sense motor                              Autocar
##                                    9                                    9
##         Maquinària d'obres i serveis                        Tractor camió
##                                    6                                    3
##                       Microbús <= 17
##                                    2
```

```
acc_old <- acc

acc[ acc$TipVeh=="Ciclomotor" | acc$TipVeh ==
      "Veh. mobilitat personal amb motor",]$TipVeh <- "Motocicleta"

acc[ acc$TipVeh=="Turisme" | acc$TipVeh=="Taxi" | acc$TipVeh==
       "Tot terreny",]$TipVeh  <- "Cotxe"

acc[ acc$TipVeh=="Autobús articulat" | acc$TipVeh=="Microbús <= 17" |
       acc$TipVeh=="Autocar" ,]$TipVeh <- "Autobús"

acc[ acc$TipVeh=="Camió rígid <= 3,5 tones" |  acc$TipVeh=="Camió rígid > 3,5 tones" |
       acc$TipVeh=="Tractor camió" | acc$TipVeh=="Desconegut"  |
       acc$TipVeh=="Altres vehicles amb motor" |  acc$TipVeh=="Camió" |
       acc$TipVeh=="Maquinària d'obres i serveis" |
       acc$TipVeh=="Tren o tramvia",]$TipVeh <- "Altres"

acc[ acc$TipVeh=="Veh. mobilitat personal sense motor" |  acc$TipVeh==
       "Altres vehicles sense motor" ,]$TipVeh <- "Bicicleta"

# Elimina filas duplicadas
acc <- acc %>% distinct()
```

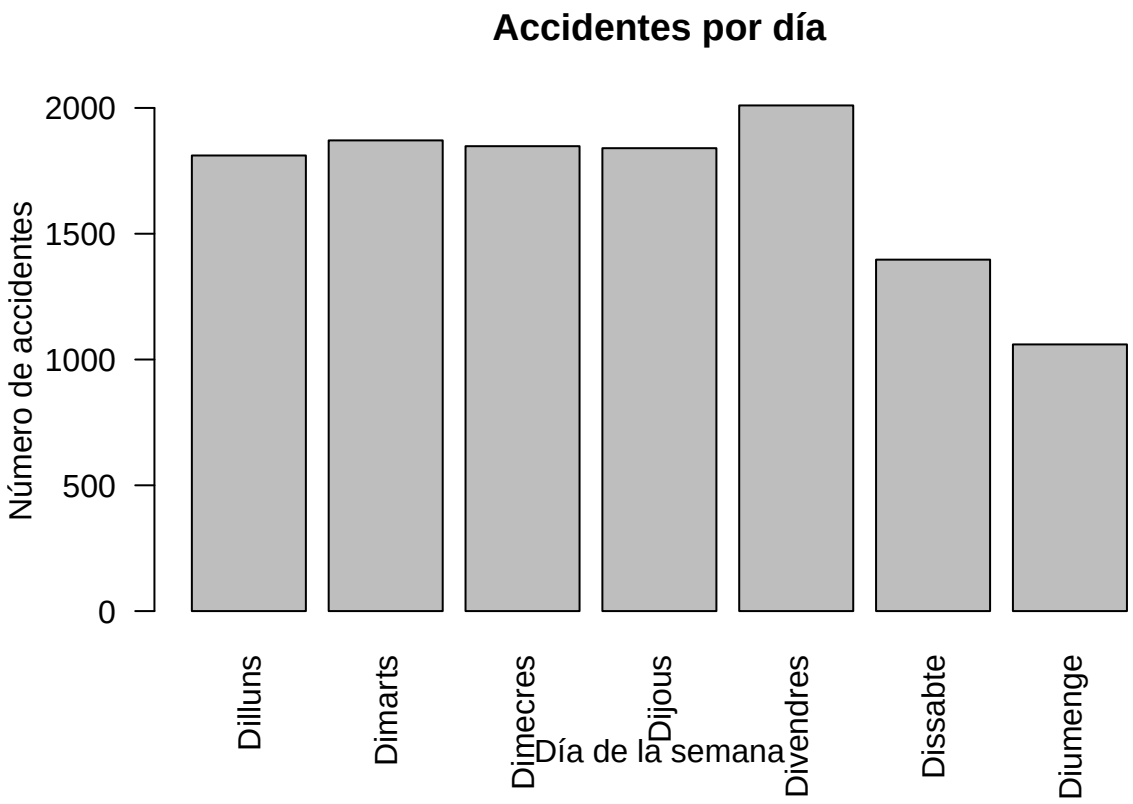
4.9 Factorizamos las variables cualitativas

```
acc$Exp <- as.factor(acc$Exp)
acc$Dist <- as.factor(acc$Dist)
acc$Dia <- as.factor(acc$Dia)
acc$Turno <- as.factor(acc$Turno)
acc$TipVeh <- as.factor(acc$TipVeh)
acc$Sexo <- as.factor(acc$Sexo)
acc$TipPer <- as.factor(acc$TipPer)

tipos$Exp <- as.factor(tipos$Exp)
tipos$TipAcc <- as.factor(tipos$TipAcc)
```

5 Análisis descriptivo

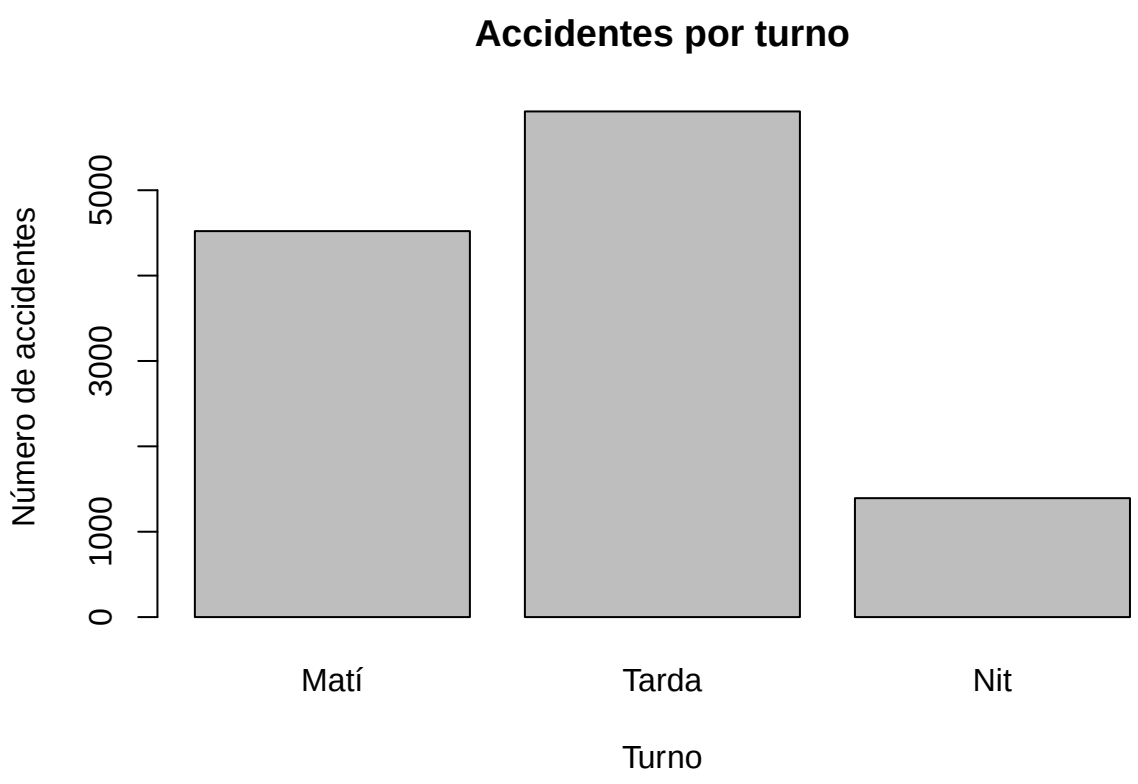
5.1 Distribución accidentes por día de la semana

```
# Se ordenan y etiquetan abreviados los niveles del factor

acc$Dia <- factor(acc$Dia, levels=c("Dilluns", "Dimarts","Dimecres","Dijous","Divendres",
                                    "Dissabte", "Diumenge"))

# Gráfico de barras
barplot(table(acc$Dia), main="Accidentes por día", xlab="Día de la semana",
        ylab="Número de accidentes", las=2)
```



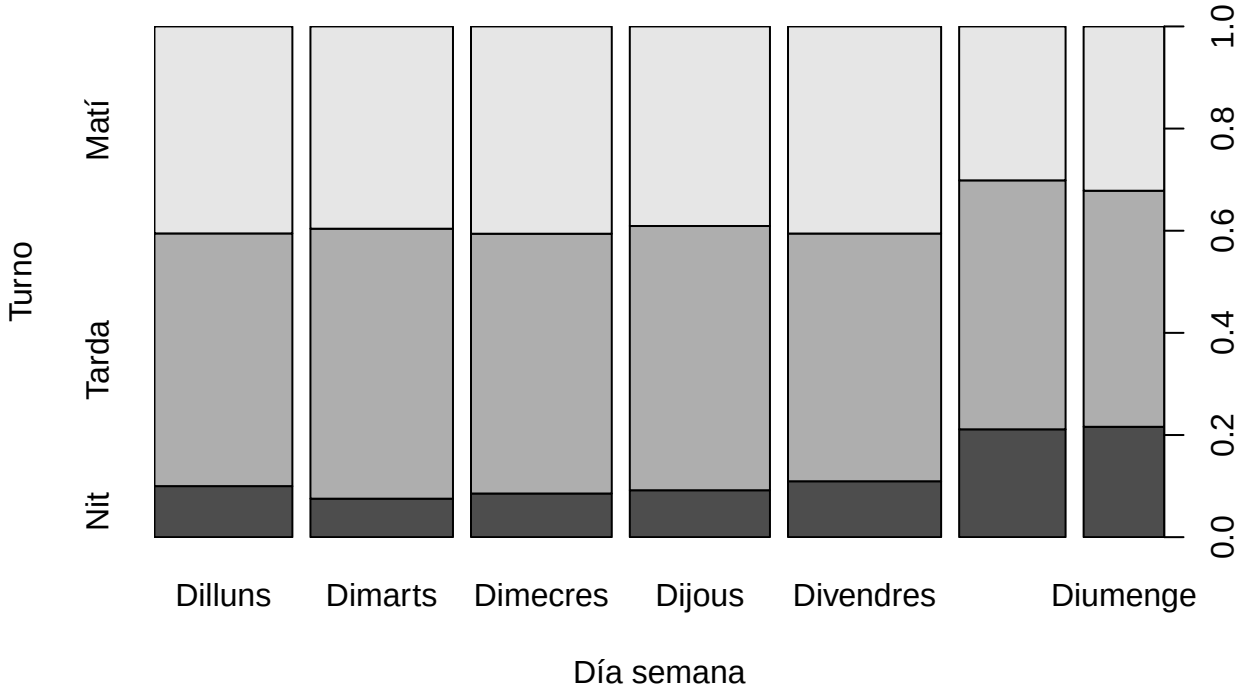
5.2 Distribución accidentes por turno

```
# Se ordenan y etiquetan abreviados los niveles del factor
acc$Turno <- factor(acc$Turno, levels=c("Matí", "Tarda", "Nit"))

# Gráfico de barras
barplot(table(acc$Turno), main="Accidentes por turno", xlab="Turno",
        ylab="Número de accidentes")
```



5.3 Distribución accidentes por turno y día de la semana

```
acc$Turno <- factor(acc$Turno, levels=c( "Nit", "Tarda","Matí"))
plot(acc$Turno ~ acc$Dia, ylab="Turno", xlab="Día semana",las=2)
```



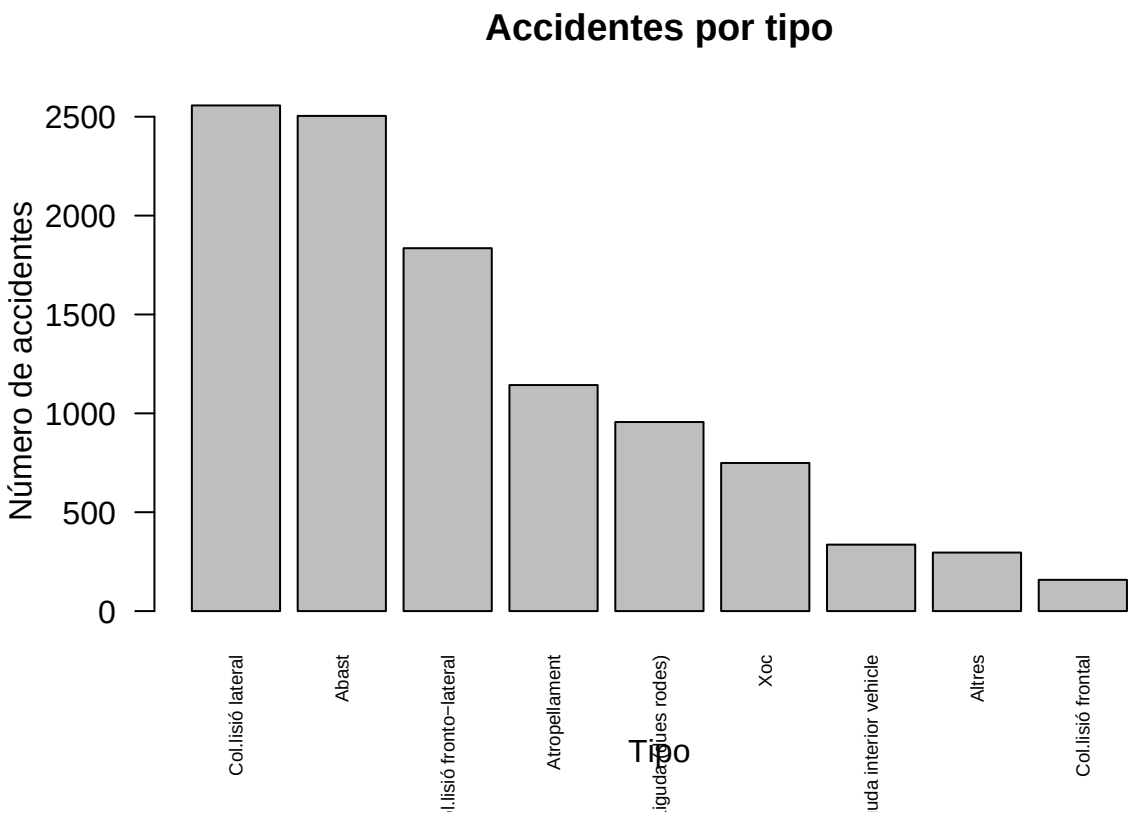
5.4 Distribución accidentes por tipo

```
kable(sort(table(tipos$TipAcc), decreasing=TRUE), col.names = c("Tipo","Número"),
      caption="Accidentes por tipo")
```

Table 1: Accidentes por tipo

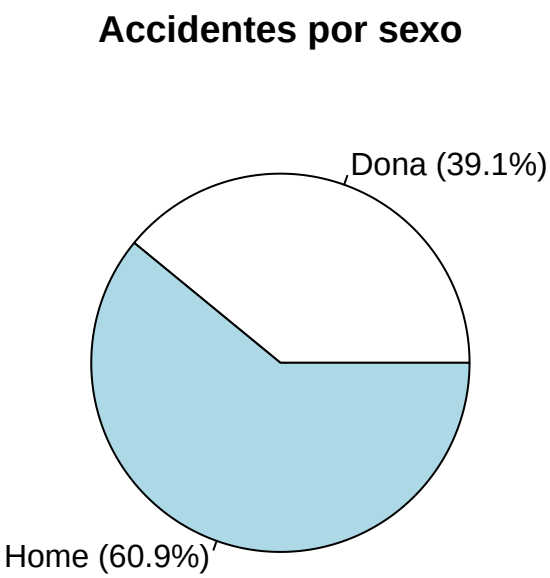
Tipo	Número
Col.lisió lateral	2557
Abast	2504
Col.lisió fronto-lateral	1835
Atropellament	1143
Caiguda (dues rodes)	956
Xoc	749
Caiguda interior vehicle	336
Altres	296
Col.lisió frontal	158

```
barplot(sort(table(tipos$TipAcc), decreasing=TRUE), main="Accidentes por tipo",
        xlab="Tipo", ylab="Número de accidentes", las=2, cex.names=0.6)
```



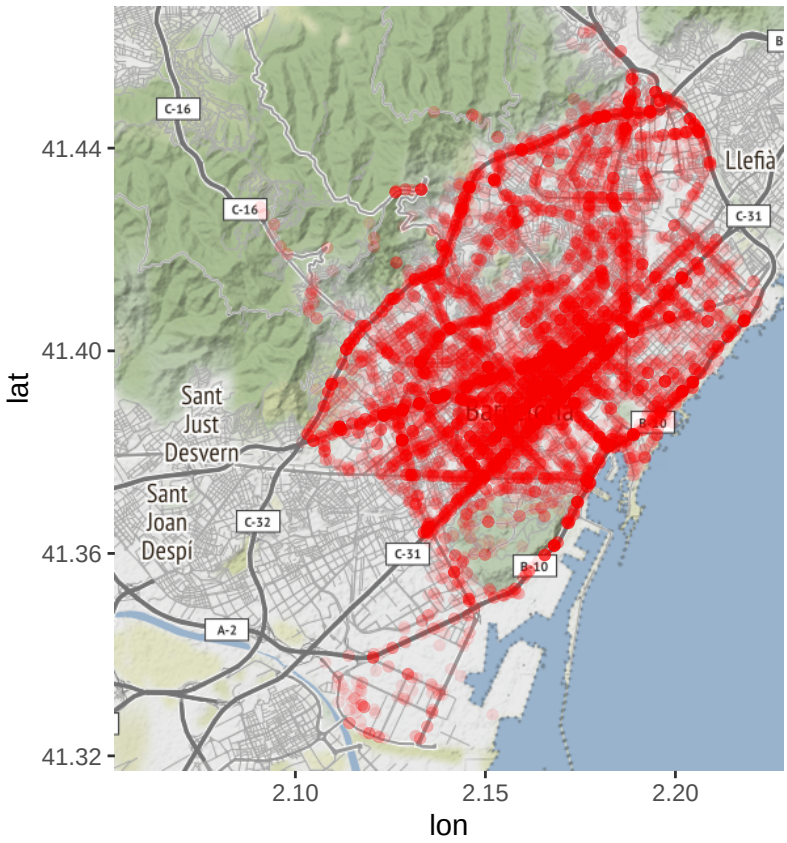
5.5 Distribución por sexos

```
# creamos una tabla con los datos de sexo
sex <- table(acc$Sexo)

#Creamos las etiquetas
labels <- sprintf("%s (%3.1f%%)", names(sex), 100*sex/sum(sex) )

pie(sex, labels, main="Accidentes por sexo")
```



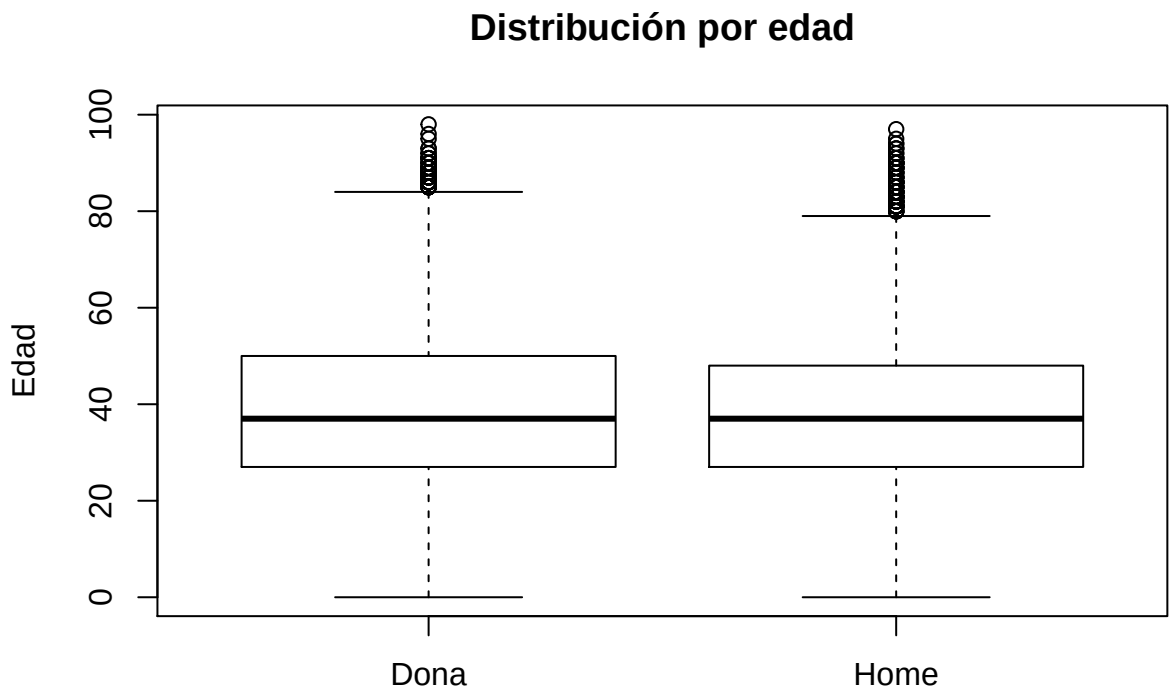
5.6 Distribución accidentes por localización

```
#Obtiene el mapa de fondo
mapa <- get_map(getbb("Barcelona"),maptype = "toner-background")

ggmap(mapa) + geom_point(aes(x = Long, y = Lat), colour="red", data = acc, alpha = .1 )
```



5.7 Distribución por edad

```
boxplot(acc$Edad~acc$Sexo, main="Distribución por edad", ylab="Edad")
```



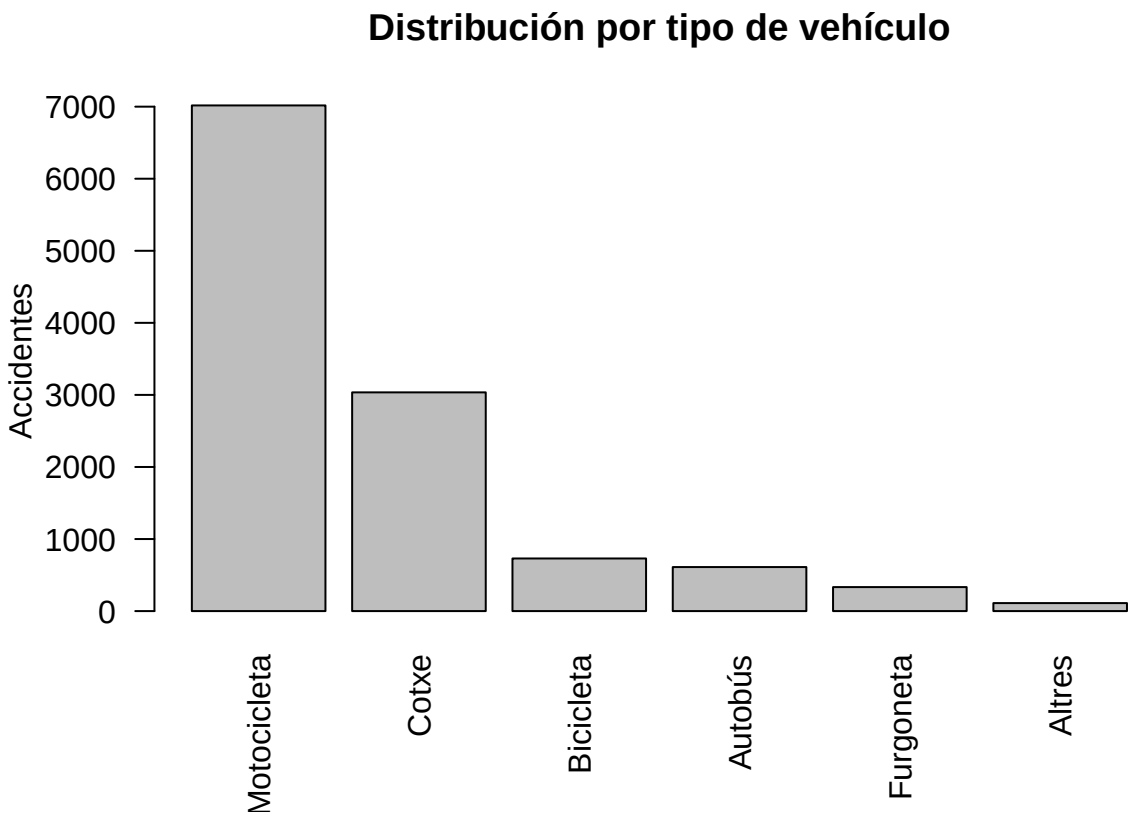
5.8 Distribución por tipo de vehiculo

```
kable(sort(table(acc$TipVeh), decreasing=TRUE), col.names = c("Tipo","Número"),
      caption="Accidentes por tipo vehiculo")
```

Table 2: Accidentes por tipo vehiculo

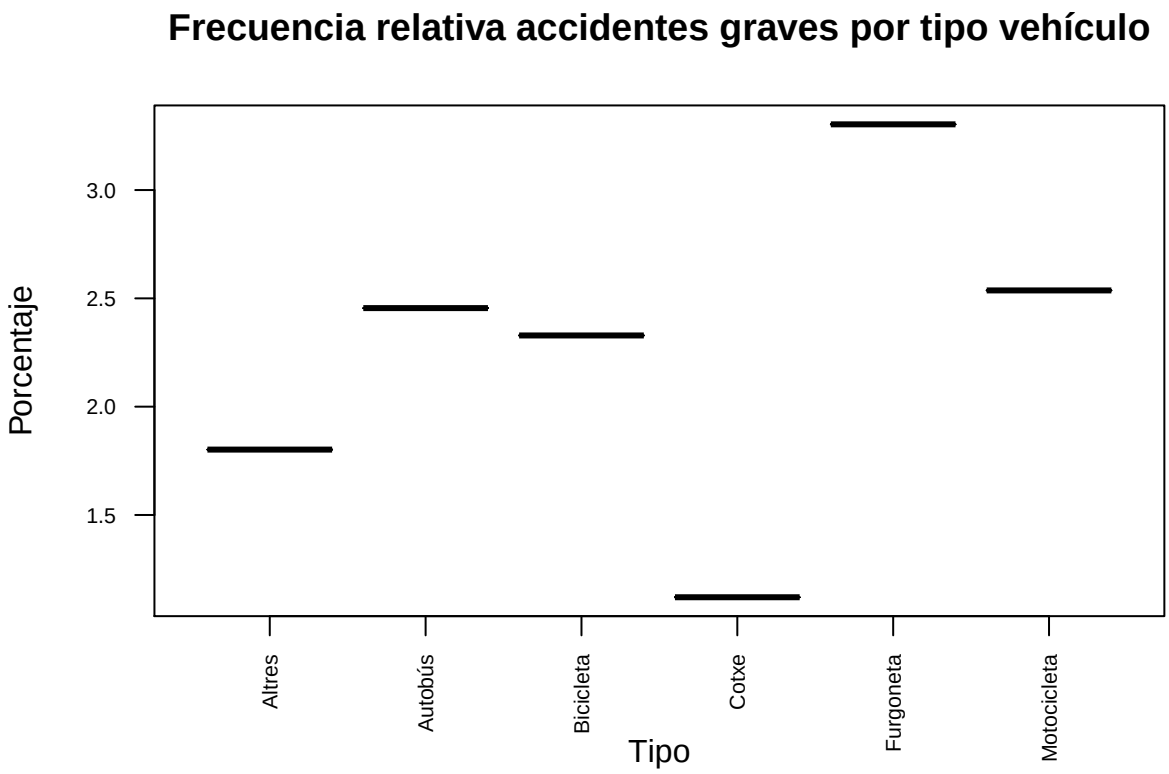
Tipo	Número
Motocicleta	7017
Cotxe	3035
Bicicleta	730
Autobús	611
Furgoneta	333
Altres	111

```
barplot(sort(table(acc$TipVeh), decreasing = TRUE), las=2,
        main="Distribución por tipo de vehículo", ylab="Accidentes")
```



5.9 Frecuencia relativa de accidentes grave por tipo de vehiculo

```
df = data.frame(table(acc$TipVeh, acc$gravedad))
pct <- group_by(df, Var1) %>% mutate(percent = (Freq/sum(Freq))*100)
plot(pct[pct$Var2==1,c(1,4)], las=2, cex.axis=0.7,
     main="Frecuencia relativa accidentes graves por tipo vehículo",
     ylab="Porcentaje", xlab="Tipo")
```



6 Análisis inferencial: ¿la edad de los accidentados graves es mayor que la de los accidentados leves?

En este punto queremos comprobar si la edad de los accidentados graves es superior a la de los accidentados leves. Para ello realizaremos un contraste de hipótesis sobre la media de edad para dos muestras independientes (el grupo de los graves y el grupo de los leves). Antes, sin embargo, vamos a comprobar si se dan los supuestos de normalidad e igualdad de varianzas en nuestra muestra.

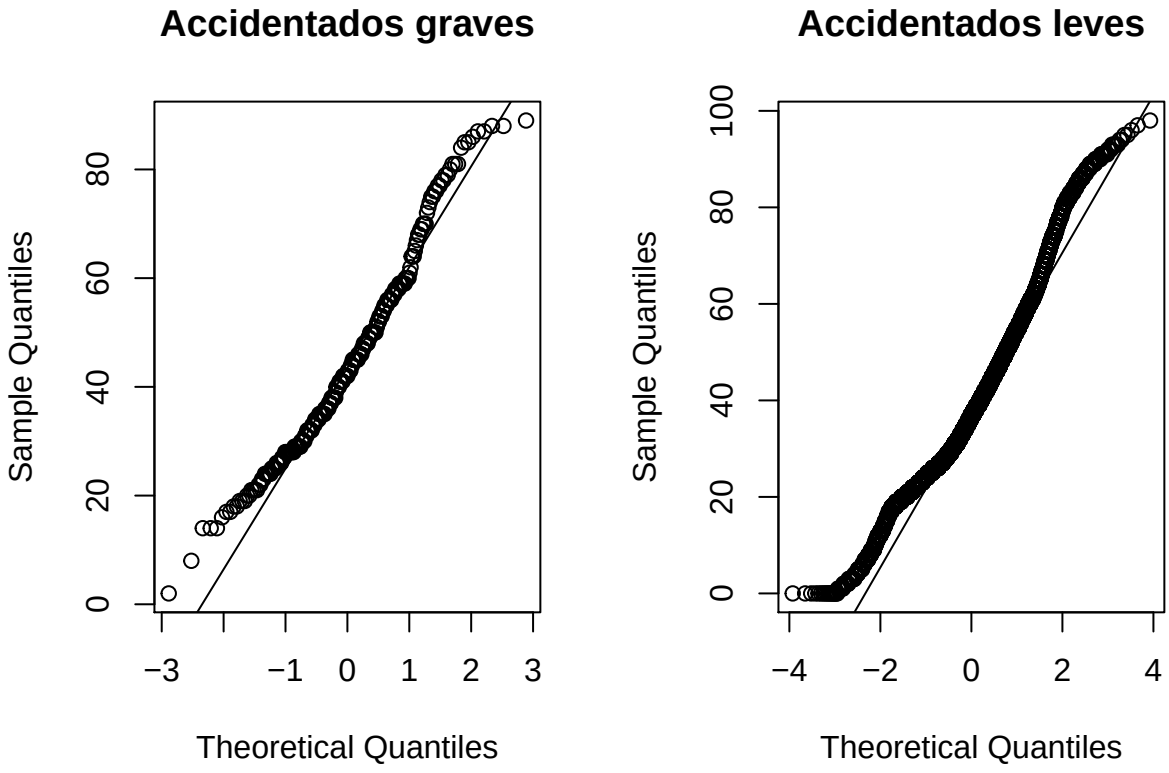
6.1 Comprobación de la normalidad

Como tenemos una muestra suficientemente grande, por el Teorema del Límite Central podríamos asumir normalidad en la distribución de la media y utilizar un test parámetrico. Aún así, vamos a examinar la normalidad de la variable Edad.

Visualmente utilizaremos el gráfico Q-Q plot y numéricamente utilizaremos el test de significancia de Shaphiro-Wilk. Este test realiza un contraste sobre la normalidad, donde la hipótesis nula es que la muestra sigue una distribución normal y la hipótesis alternativa es que no la sigue.

```
acc.graves <- acc[acc$gravedad==1,]$Edad
acc.nograves <- acc[acc$gravedad==0,]$Edad

# Gráfico Q-Q para la muestra de accidentados graves
par( mfrow=c(1,2))
```

```r
qqnorm( acc.graves, main="Accidentados graves" )
qqline( acc.graves ) # la recta

# Gráfico Q-Q para la muestra de accidentados no graves
qqnorm( acc.nograves, main="Accidentados leves" )
qqline( acc.nograves )
```




```r
shapiro.test (acc.graves )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  acc.graves
## W = 0.97099, p-value = 4.343e-05
```

```r
# El test no admite más de 5000 elementos
shapiro.test ( sample(acc.nograves,5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(acc.nograves, 5000)
## W = 0.96759, p-value < 2.2e-16
```

Visualmente podemos apreciar que la distribución de la variable Edad no se ajusta a la normal. El test de Shaphiro-Wilk corrobora esa impresión, ya que en ambos casos nos proporciona un p-valor muy inferior al nivel de significancia (0.05).

6.2 Comprobación de la igualdad de varianzas

A continuación, comprobamos la homegeneidad de las varianzas con la prueba F, que realiza un contraste donde la hipótesis nula es que las varianzas sean iguales.

```
var.test(acc.graves, acc.nograves, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  acc.graves and acc.nograves
## F = 1.246, num df = 256, denom df = 11579, p-value = 0.00992
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.053311 1.496138
## sample estimates:
## ratio of variances
##           1.246028
```

El test nos proporciona un p-valor inferior al nivel de significación, por lo que podemos rechazar la hipótesis nula. Tenemos por tanto varianzas significativamente distintas en ambas muestras.

6.3 Contraste de hipótesis

Puesto que no tenemos varianzas homegeneas, para este contraste utilizaremos el test Welch. En R se puede especificar el parámetro var.equal=FALSE para indicar que se realice este test.

Las hipótesis serán:

H0: media_graves = media_nograves

Ha: media_graves > media_nograves

```
t.test(acc.graves, acc.nograves, var.equal = FALSE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  acc.graves and acc.nograves
## t = 5.0461, df = 265.2, p-value = 4.191e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.819978      Inf
## sample estimates:
## mean of x mean of y
##  44.63035  38.95337
```

El test arroja un p-valor muy pequeño, por tanto podemos rechazar la hipotesis nula (las medias de edad son iguales) y tomar la hipótesis alternativa: la media de edad de los accidentados graves es mayor que la de los accidentados leves.

7 ¿Cuáles son los factores de riesgo que determinan la gravedad de un accidente?

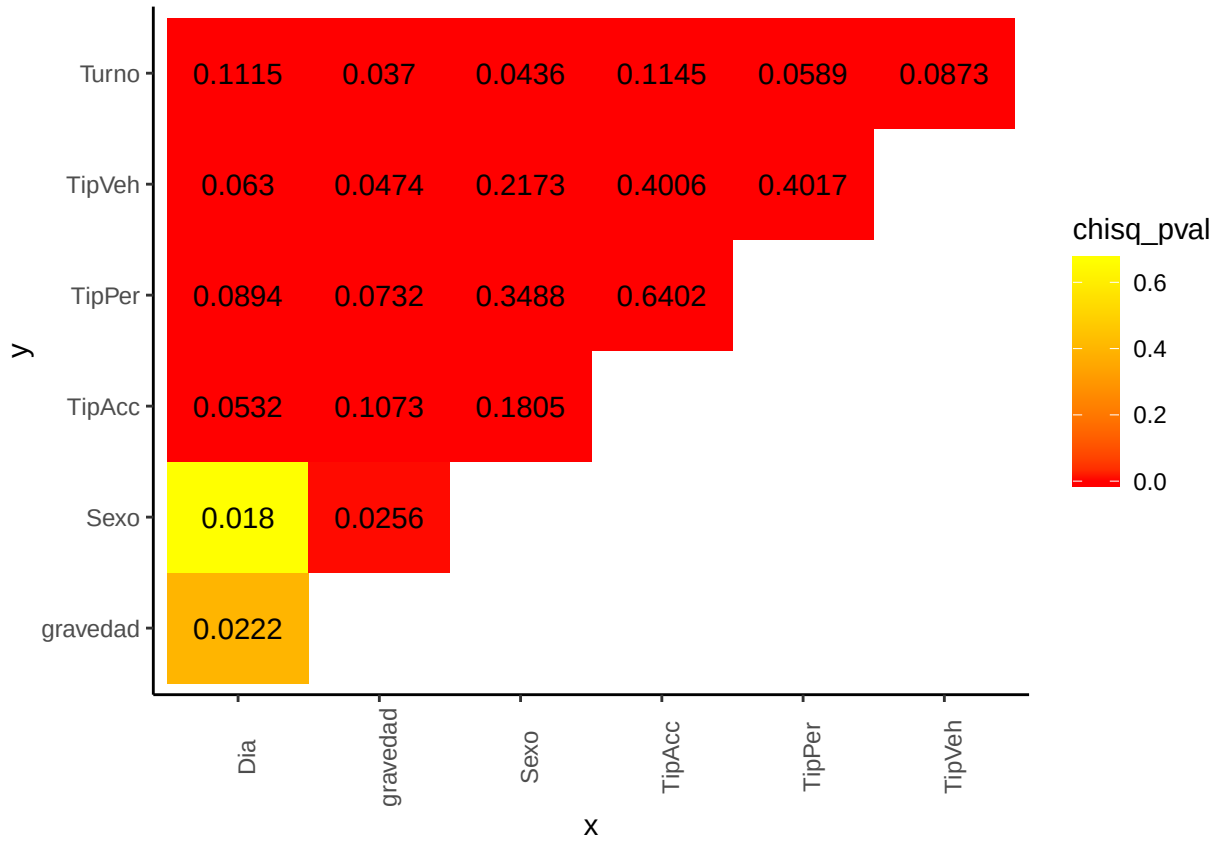
En este apartado trataremos de identificar aquellos factores que influyen en mayor medida en la gravedad de un accidente. Para ello, plantearemos un modelo de regresión generalizado donde la variable respuesta será Gravedad, que recordemos es una variable dicotómica (0=accidentado leve, 1=accidentado grave) y las variables explicativas serán Edad, Sexo, Tipo de accidente, Tipo de vehículo, Dia semana, Turno (mañana, tarde, noche) y Tipo de persona (conductor, acompañante, viandante).

7.1 Análisis de correlación

En primer lugar vamos a analizar la independencia de las variables cualitativas. Entre las variables predictoras, nos puede interesar no incluir en el modelo aquellas que estén muy correlacionadas entre sí. COn respecto a la variable respuesta, nos puede ayudar a conocer qué variables influyen más en el resultado.

Utilizaremos el test chi-cuadrado, que mide la correlación entre dos variables cualitativas. Se presenta un gráfico donde el color depende del p-valor y la cifra en cada recuadro se corresponde con el valor v de Cramers. Por un lado, si el p-valor es inferior al nivel de significancia indicará que no hay una relación estadísticamente significativa entre esas variables (son independientes). Por otro lado, el valor V es un índice que mide la fuerza de la asociación y se sitúa entre 0=no hay asociación y 1=totalmente asociadas.

```
# Incluimos el tipo de accidente en el dataset
acc <- merge(acc, tipos)

cualitativas <- c("gravedad","Dia","Turno","TipVeh","TipPer", "Sexo", "TipAcc")

# Función para obtener el p-valor de chi-cuadrado y la V de Cramers
chi_square = function(x,y, df) {
    tbl = acc %>% select(x,y) %>% table()
    chisq_pval = round(chisq.test(tbl)$p.value, 4)
    cramV = round(cramersV(tbl), 4)
    data.frame(x, y, chisq_pval, cramV) }

# Crea combinaciones únicas de las columnas
df_comb = data.frame(t(combn(sort(cualitativas), 2)), stringsAsFactors = F)

# Aplica la función chi_square a cada combinación de variables
df_res = map2_df(df_comb$X1, df_comb$X2, chi_square)

# Gráfico de resultados
df_res %>%
  ggplot(aes(x,y,fill=chisq_pval))+
  geom_tile()+
  geom_text(aes(x,y,label=cramV))+
  scale_fill_gradient(low="red", high="yellow")+
  theme_classic()+
  theme(axis.text.x = element_text(angle = 90))
```



7.2 Modelos de regresión logística

A continuación se crearán varios modelos de regresión logística con distintas combinaciones de las variables predictoras. Utilizaremos un conjunto de datos de entrenamiento y otro de test, en una proporción 80/20. Evaluaremos la calidad de los modelos a través de varias medidas:

- Una de las más utilizadas es el **AIC** (Criterio de información de Akaike), que se puede utilizar para comparar la calidad de varios modelos con la misma variable dependiente: cuanto más pequeño es este índice, mejor es la calidad del modelo.
- Para describir el rendimiento del modelo de clasificación contra un conjunto de test, utilizaremos una matriz de confusión. En la diagonal principal de esta matriz se muestran los verdaderos positivos y los verdaderos negativos, es decir los casos en los que la predicción ha sido correcta. En la otra diagonal se muestran los falsos positivos (el valor real era No y el sistema ha predicho SÍ) y los falsos negativos (el valor real era SÍ y el sistema ha predicho NO). A partir de aquí podemos calcular fácilmente la **precisión** del modelo, es decir el porcentaje de veces que acierta en la predicción.
- Por último, también se realizará un gráfico de la **curva ROC** y se calculará el valor **AUC** (area under the curve) que son medidas típicas de rendimiento para los clasificadores binarios. La curva ROC viene dada por el ratio de Verdaderos Positivos contra Falsos Positivos en distintos umbrales. El valor AUC es el área debajo de la curva ROC. Como regla general, un modelo con buena capacidad predictiva debe tener un AUC más cercano a 1 (1 es ideal) que a 0.5.

Antes de construir los conjuntos de entrenamiento y test tendremos que atender a otra cuestión: en nuestro caso las muestras (graves y no graves) están muy desbalanceadas (una tiene un tamaño muy inferior a la

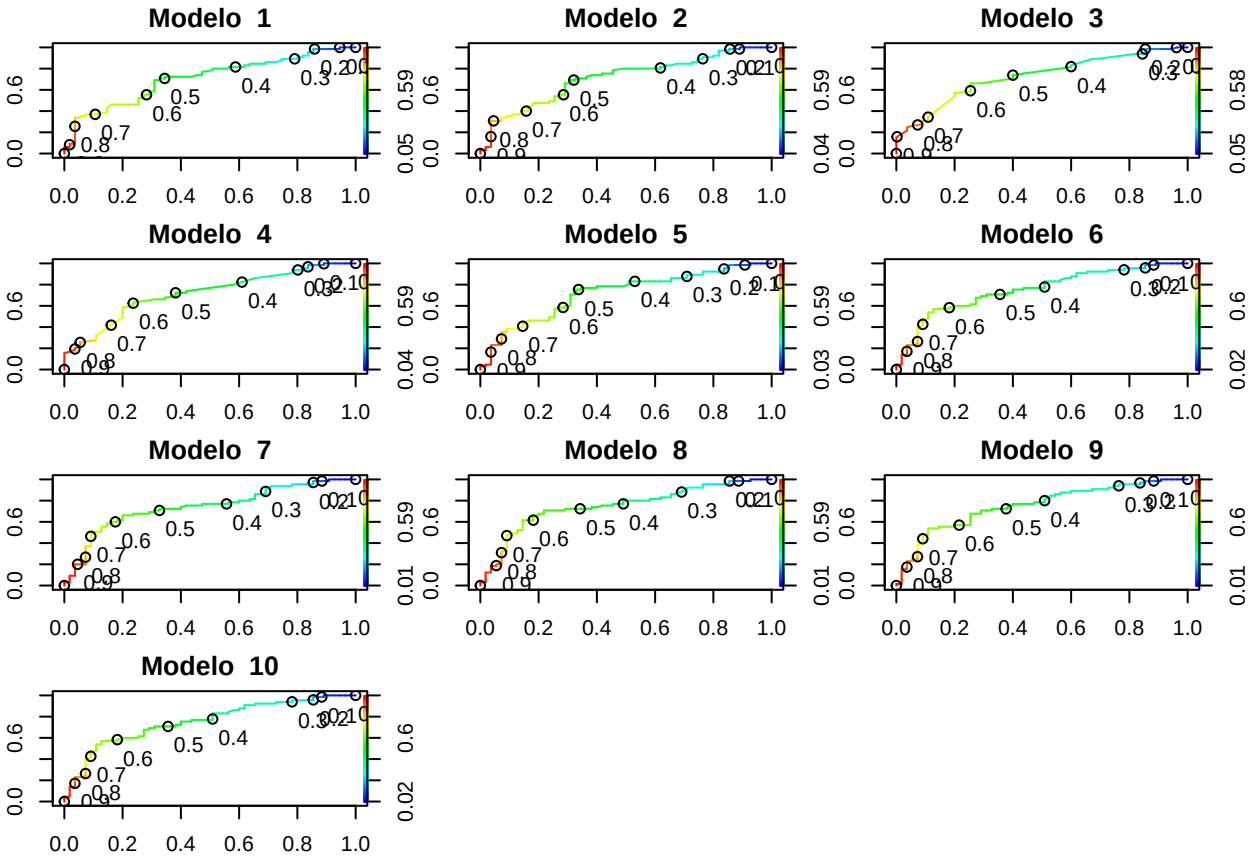
otra). Esto provoca sesgo en el modelo, que acaba tendiendo hacia el lado donde la muestra es mayor. Por ello realizaremos una disminución del conjunto más grande, para equiparar el tamaño de ambas muestras; este proceso se conoce como oversampling. Se pierde información, pero aumenta la precisión del modelo.

```
set.seed(123)


# Establecemos el nivel de referencia en los predictores con múltiples niveles
# que serán convertidos en dummies

acc$TipVeh <- relevel(acc$TipVeh, ref = "Cotxe")
acc$TipPer <- relevel(acc$TipPer, ref = "Conductor")
acc$Dia<- relevel(acc$Dia, ref = "Dilluns")
acc$Turno <- relevel(acc$Turno, ref = "Matí")
acc$Sexo <- relevel(acc$Sexo, ref = "Dona")

# Oversampling

graves <- acc[acc$gravedad==1,]
nograves <- acc[sample(nrow(acc[acc$gravedad==0,]), nrow(graves), replace=FALSE),]
accsampled <- rbind(graves, nograves)


# Creación conjunto de entrenamiento y test con una proporcion 80/20

ind <- sample(2, nrow(accsampled), replace=TRUE, prob=c(0.8, 0.2))
train <- accsampled[ind==1,]
test  <- accsampled[ind==2,]


# Creación de los modelos

glm_model1 <- glm(gravedad ~  Edad + TipAcc + TipVeh, data=train, family=binomial)

glm_model2 <- glm(gravedad ~  Edad + TipAcc + TipVeh + TipPer, data=train,
                  family=binomial)

glm_model3 <- glm(gravedad ~  Sexo + TipAcc + TipVeh, data=train, family=binomial)

glm_model4 <- glm(gravedad ~  Sexo + TipAcc + TipVeh + TipPer, data=train,
                  family=binomial)

glm_model5 <- glm(gravedad ~  Edad + Sexo + Dia + TipAcc + TipVeh + TipPer, data=train,
                  family=binomial)

glm_model6 <- glm(gravedad ~  Edad + Turno + TipAcc + TipVeh + TipPer, data=train,
                  family=binomial)

glm_model7 <- glm(gravedad ~  Edad + Dia + TipAcc + TipVeh + TipPer + Turno, data=train,
                  family=binomial)

glm_model8 <- glm(gravedad ~  Edad + Sexo + Dia + TipAcc + TipVeh + TipPer + Turno,
                  data=train, family=binomial)
```

```r
glm_model9 <- glm(gravedad ~  Edad + Sexo + Turno + TipAcc +  TipVeh + TipPer ,
                  data=train, family=binomial)

glm_model10 <- glm(gravedad ~  Edad + Turno + TipAcc + TipVeh + TipPer, data=train,
                   family=binomial)

models <- list(glm_model1, glm_model2, glm_model3, glm_model4, glm_model5, glm_model6,
               glm_model7, glm_model8, glm_model9, glm_model10)


# Comprobar AIC, precisión, AUC y curva ROC de cada modelo
par(mar = rep(2, 4))
par( mfrow=c(4,3))

for (i in 1:10) {

  # Crea la matriz de confusión
  predictTest = predict(models[[i]], type = "response", newdata = test)
  table_mat <- table(test$gravedad, predictTest>0.5)

  # Calcula la precisión a partir de la matriz de confusión
  accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)

  # Curva ROC y cálculo de AUC (Area Under Curve)
  ROCRpred <- prediction(predictTest, test$gravedad)
  ROCRperf <- performance(ROCRpred, measure = "tpr", x.measure = "fpr")

  auc <- performance(ROCRpred, measure = "auc")
  auc <- auc@y.values[[1]]

  cat("Modelo ",i," AIC:",models[[i]]$aic," Precisión: ",
      accuracy_Test, " AUC: ", auc,"\n")

  # Gráfico curva ROC
  plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at
       = seq(0,1,0.1), main=paste("Modelo ", as.character(i), sep=" "))

}
```

```
## Modelo  1  AIC: 581.4057  Precisión:  0.6916667  AUC:  0.708951

## Modelo  2  AIC: 576.6028  Precisión:  0.6916667  AUC:  0.7109091

## Modelo  3  AIC: 591.8514  Precisión:  0.6666667  AUC:  0.7261538

## Modelo  4  AIC: 586.2973  Precisión:  0.6666667  AUC:  0.7183217

## Modelo  5  AIC: 582.6692  Precisión:  0.7166667  AUC:  0.7195804

## Modelo  6  AIC: 565.1218  Precisión:  0.6833333  AUC:  0.7541259

## Modelo  7  AIC: 570.1957  Precisión:  0.7  AUC:  0.7439161
```

```
## Modelo  8  AIC: 571.4665  Precisión:  0.7  AUC:  0.7509091

## Modelo  9  AIC: 566.5034  Precisión:  0.6833333  AUC:  0.7541259

## Modelo  10  AIC: 565.1218  Precisión:  0.6833333  AUC:  0.7541259
```



7.3 Interpretación del modelo y odds ratios

```r
# Tomamos el mejor modelo
modelo <- models[[6]]

summary(modelo)
```

```
##
## Call:
## glm(formula = gravedad ~ Edad + Turno + TipAcc + TipVeh + TipPer,
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2924  -0.9491   0.2748   0.9351   2.1970
##
## Coefficients:
```

```
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -3.954283   0.564641  -7.003 2.50e-12 ***
## Edad                           0.027742   0.007211   3.847 0.000120 ***
## TurnoNit                       1.334950   0.369415   3.614 0.000302 ***
## TurnoTarda                     0.691560   0.247070   2.799 0.005125 **
## TipAccAltres                   1.322075   0.860150   1.537 0.124287
## TipAccAtropellament            0.905796   0.609257   1.487 0.137088
## TipAccCaiguda (dues rodes)     0.745060   0.454137   1.641 0.100879
## TipAccCaiguda interior vehicle 3.030135   0.873746   3.468 0.000524 ***
## TipAccCol.lisió frontal        1.209964   0.614301   1.970 0.048877 *
## TipAccCol.lisió fronto-lateral 1.221445   0.348165   3.508 0.000451 ***
## TipAccCol.lisió lateral        0.217892   0.362822   0.601 0.548140
## TipAccXoc                      3.025668   0.741534   4.080 4.50e-05 ***
## TipVehAltres                  -0.723665   1.498890  -0.483 0.629237
## TipVehAutobús                 -0.527700   0.791740  -0.667 0.505088
## TipVehBicicleta                1.284651   0.511384   2.512 0.012001 *
## TipVehFurgoneta                0.202336   0.650018   0.311 0.755589
## TipVehMotocicleta              1.787187   0.338331   5.282 1.28e-07 ***
## TipPerPassatger               -0.649793   0.371977  -1.747 0.080661 .
## TipPerVianant                  1.929309   0.639037   3.019 0.002535 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 659.57  on 475  degrees of freedom
## Residual deviance: 527.12  on 457  degrees of freedom
## AIC: 565.12
##
## Number of Fisher Scoring iterations: 4
```

En la columna Pr(>|z|) encontramos los p-valores que indican la significancia de cada variable. Un valor por debajo del nivel de significancia (por defecto 0.05) indica que existe asociación entre esa variable y el resultado.

En la columna "Estimate" encontramos los coeficientes que nos indican la fuerza y dirección de la relación. Un número cercano a 0 indica poca influencia. Si es positivo, indica una influencia positiva, es decir, en presencia de la variable aumenta la probabilidad del resultado que estamos estudiando. Esto hay que interpretarlo siempre respecto a los valores de referencia que hemos establecido.

Por ejemplo, vemos que la variable "TipVehMotocicleta" es significativa (tiene un p-valor muy pequeño) y un coeficiente de 1.7, lo que indica que el hecho de tener un accidente en motocicleta aumenta el riesgo de sufrir heridas graves (respecto a ir en coche, que es la categoría base).

Si elevamos ese coeficiente al numero "e" obtenemos el odd ratio, que ayuda a explicar mejor esta relación.

```
exp(modelo$coefficients["TipVehMotocicleta"])
```

```
## TipVehMotocicleta
##          5.972629
```

El odd ratio es el cociente entre la probabilidad de que ocurra un evento dividido por la probabilidad de que no ocurra. Toma valores entre 0 e infinito: si el OR < 1 indicará una asociación negativa entre las variables y si OR > 1 entonces señala una asociación positiva. Si oR $= 1$ implica que no hay asociación entre las variables.

En nuestro caso, el odd ratio entre "Motocicleta" y "Accidente grave" es de 5.9, lo que se puede interpretar como que el riesgo de sufrir heridas graves (o la muerte) cuando se tiene un accidente en motocicleta es 5.9 veces más que cuando se tiene el mismo accidente en coche (que es la categoría de referencia). Por "mismo accidente" se entiende que el resto de variables predictoras de nuestro modelo se mantienen iguales.

Podemos proceder igual con otras variables significativas:

```
exp(modelo$coefficients["TurnoNit"])
```

```
## TurnoNit
## 3.799807
```

```
exp(modelo$coefficients["TurnoTarda"])
```

```
## TurnoTarda
##   1.996828
```

```
exp(modelo$coefficients["TipAccCol.lisió frontal"])
```

```
## TipAccCol.lisió frontal
##               3.353364
```

```
exp(modelo$coefficients["TipAccCol.lisió fronto-lateral"])
```

```
## TipAccCol.lisió fronto-lateral
##                       3.392087
```

```
exp(modelo$coefficients["TipAccXoc"])
```

```
## TipAccXoc
##  20.60776
```

```
exp(modelo$coefficients["TipVehBicicleta"])
```

```
## TipVehBicicleta
##        3.613406
```

```
exp(modelo$coefficients["TipPerVianant"])
```

```
## TipPerVianant
##      6.884753
```

El OR entre TurnoNit y Accidente Grave es de 3.79, lo que indica que la posibilidad de que el desenlace de un accidente sea grave por la noche es 3.79 veces más que durante la mañana (categoría base), manteniéndose iguales el resto de variables. No hay que olvidar que este estudio incluye sólo unas pocas variables, y que los factores que pueden explicar la gravedad del accidente pueden ser muchos más: la meteorología, la ubicación, ...

7.4 Predicción

A partir del modelo anterior, podemos tratar de hacer algunas predicciones sobre la probabilidad de que se sufra un accidente con lesiones graves o muerte en varias circunstacias.

Por ejemplo, según este modelo, ¿cuál será la probabilidad de que un conductor de 70 años, que viaja en coche como conductor, por la mañana, y sufre un accidente con una colisión fronto-lateral, sufra heridas graves o muera?

```
# Valores de los predictores
predictdata = data.frame(Edad=70, Turno="Matí", TipAcc="Col.lisió fronto-lateral",
                         TipVeh="Cotxe", TipPer="Conductor")

# Predicción
predict(modelo, predictdata, type = "response")
```

```
##         1
## 0.3119723
```

La probabilidad de sufrir un accidente grave en este caso es del 31.1%

Si en lugar de conducir por la mañana, este mismo accidente se produce por la noche, ¿cuál sería la probabilidad?

```
# Valores de los predictores
predictdata = data.frame(Edad=70, Turno="Nit", TipAcc="Col.lisió fronto-lateral",
                         TipVeh="Cotxe", TipPer="Conductor")

# Predicción
predict(modelo, predictdata, type = "response")
```

```
##         1
## 0.6327507
```

En este caso, la probabilidad aumenta hasta un 63.27%.

Si en lugar de con coche, circulase con una motocicleta, ¿cuál sería la probabilidad?

```
# Valores de los predictores
predictdata = data.frame(Edad=70, Turno="Nit", TipAcc="Col.lisió fronto-lateral",
                         TipVeh="Motocicleta", TipPer="Conductor")

# Predicción
predict(modelo, predictdata, type = "response")
```

```
##         1
## 0.9114301
```

En este caso, se alcanza una probabilidad del 91,14% de sufrir lesiones graves o muerte.

Si el conductor, en lugar de 70 años, tuviera 20, en este último caso, cuál sería la probabilidad de un accidente grave?

```r
# Valores de los predictores
predictdata = data.frame(Edad=20, Turno="Nit", TipAcc="Col.lisió fronto-lateral",
                         TipVeh="Motocicleta", TipPer="Conductor")

# Predicción
predict(modelo, predictdata, type = "response")
```

```
##         1
## 0.7199342
```

En este caso, la probabilidad baja al 71,9%.

7.5 Escribir ficheros de salida

```r
write.csv(acc, file="Accidentados2018.csv")
write.csv(accsampled, file="Accidentados2018_predict_graves_no_graves.csv")
```

8 Conclusiones

En base al estudio anterior podemos obtener algunas conclusiones que nos ayudan a caracterizar los accidentes de tráfico en la ciudad de Barcelona y determinar los factores de riesgo en los mismos.

- Aproximadamente la mitad de los accidentes se producen durante la tarde.
- El viernes es el día de mayor concentración de accidentes.
- Los fines de semana se producen menos accidentes que durante los días laborales; sin embargo, en estos dos días se produce un porcentaje mayor de accidentes durante la noche.
- El tipo más frecuente de accidente se produce por una colisión lateral, seguido del alcance (por detrás), la colisión fronto-lateral, el atropello y la caída en vehículos de 2 ruedas.
- La mayoría de accidentes se concentran en el centro de la ciudad (zona del Eixample), en las rondas y en las principales vías (Diagonal y Gran Vía)
- Los accidentados en motocicleta (7017) doblan a los accidentados que viajaban en coche (3035). Les siguen los accidentados en bicicleta (730).
- Respecto a la gravedad, se comprueba que la edad media de los accidentados graves es superior a los accidentados leves.
- Respecto al tipo de accidente, los choques contra objetos estáticos o por salída de la vía, son los que producen los accidentes más graves, seguidos de las caídas en el interior de los autobuses o autocares y las colisiones frontales o fronto-laterales.
- Los vehículos más inseguros son las motocicletas y las bicicletas. Por ejemplo, la posibilidad de resultar herido grave o muerto en un accidente de motocicleta es casi 6 veces mayor que si se tuviese el msimo accidente en coche.