UOC - Tipología y Ciclo de vida de los datos

PRAC2: Limpieza y análisis de datos

Víctor María Cardoner Álvarez
José Oriol Bielsa Nogaledo

Índice

1. Descripción del dataset

El dataset que vamos a usar se llama “Accidents” (*accidents_2017.csv*) y pertenece al portal de Open Data de Barcelona que se ofrece en kaggle.com. Su dirección exacta es: https://www.kaggle.com/xvivancos/barcelona-data-sets.
Este conjunto de datos corresponde al listado de accidentes registrados por la Policía Local de Barcelona en 2017, obtenido del portal **Open Data BCN**.

Los campos del conjunto de datos son:

Campo	Descripción	Valores
Id	Número indicativo del expediente del accidente	
District.Name	Nombre del distrito donde se ha producido el accidente	Distritos de Barcelona, 11 valores
Neighborhood.Name	Nombre del barrio donde se ha producido el accidente	Barrios de Barcelona, 74 valores
Street	Nombre de la calle donde se ha producido el accidente	4253 calles diferentes
Weekday	Día de la semana en que se ha producido el accidente	"Monday" hasta "Sunday"
Month	Mes en que se ha producido el accidente	"January" hasta "December"
Day	Día del mes en que se ha producido el accidente	1 a 31
Hour	Hora a la que se producido el accidente	0 a 23
Part.of.the.day	Momento del día en que se ha producido el accidente	"Morning", "Afternoon", "Night"
Mild.injuries	Número de heridos leves	0 a 10
Serious.injuries	Números de heridos graves	0, 1, 2 o 4
Victims	Víctimas totales	0 a 10
Vehicles.involved	Vehículos involucrados en el accidente	0 a 14
Longitude/Latitude	Coordenadas GPS del accidente	

En el proyecto RMD existe un estudio detallado de todos los atributos del dataset, en caso de querer ver mayor detalle.

La información de este conjunto de datos nos permite estudiar los datos de accidentes en **tres ejes** de datos:
(1) Accidentes por **ubicación geográfica**, **(2)** por **ubicación temporal**, y **(3)** por **gravedad** del accidente en términos de vehículos implicados y heridos.

Visto esto, existen no una, sino **varias preguntas** que podemos tratar de resolver con este dataset:

1. **Identificar puntos negros en la ciudad**
   Esto nos permite identificar si existen focos de problemas en la circulación, de manera que se pueda estudiar si existen medidas para solucionarlo.
2. **Identificar relación entre Meses/Días con mayor siniestralidad**
   Detectar alguna tendencia o anormalidad en cuanto a distribución de la siniestralidad registrada en Barcelona.
3. **Identificar si existe mayor siniestralidad a principios o finales de mes**
   Los principios y finales de mes tienen ciertas particularidades a nivel social, que pueden provocar que exista o no una cierta tendencia.
4. **Identificar si existe correlación entre heridos graves en función de horario nocturno o diurno**
   De nuevo, el tipo de tráfico y hábitos difieren mucho en horario diurno y nocturno, vale la pena estudiar si esto muestra alguna evidencia estadística en el dataset.
5. **Identificar si existe relación entre accidentes con múltiples coches implicados con el hecho de ser o no fin de semana**
   Aplicaría el mismo comentario que el punto anterior.
6. **Identificar si existe mayor siniestralidad en periodo vacacional**
   Aplicaría el mismo comentario que el punto anterior.
7. **Identificar franjas horarias con mayor siniestralidad**
   Es posible que a determinadas horas se produzcan picos de accidentes, pero es interesante estudiar si existe alguna tendencia que se pueda confirmar estadísticamente.

2. Integración y selección de los datos de interés a analizar

Referente a la integración de datos, en el caso que nos ocupa entendemos que no aplicaría realizar ningún proceso ni transformación particular. Simplemente vamos a leer el dataset, y analizar las características básicas.
Observamos que tenemos un dataset de **10.339 registros**; con **15 atributos** de tipo categóricos y numéricos. Mostramos las primeras 10 filas, para hacernos una primera idea de los datos a grandes rasgos:

	Id <fctr>	District.Name <fctr>	Neighborhood.Name <fctr>	Street <fctr>	Weekday <fctr>	Month <fctr>	Day <int>	Hour <int>	Part.of.the.day <fctr>
10334	2017S005030	Sant Andreu	el Bon Pastor	Litoral (Besòs)	Thursday	June	8	19	Afternoon
10335	2017S003667	Sant Andreu	el Bon Pastor	Litoral (Llobregat)	Tuesday	April	25	8	Morning
10336	2017S001896	Sant Andreu	el Bon Pastor	PL MONTERREY	Wednesday	March	8	12	Morning
10337	2017S010718	Sant Andreu	el Bon Pastor	Litoral (Llobregat)	Thursday	December	28	8	Morning
10338	2017S006145	Sant Andreu	el Bon Pastor	Litoral (Besòs)	Friday	July	14	14	Afternoon
10339	2017S000178	Sant Andreu	el Bon Pastor	CIUTAT D'ASUNCIÓN	Sunday	January	8	20	Afternoon

6 rows | 1-10 of 15 columns

Weekday <fctr>	Month <fctr>	Day <int>	Hour <int>	Part.of.the.day <fctr>	Mild.injuries <int>	Serious.injuries <int>	Victims <int>	Vehicles.involved <int>	Longitude <dbl>	Latitude <dbl>
Thursday	June	8	19	Afternoon	1	0	1	2	2.205163	41.44476
Tuesday	April	25	8	Morning	1	0	1	3	2.201800	41.39200
Wednesday	March	8	12	Morning	1	0	1	2	2.206013	41.44344
Thursday	December	28	8	Morning	1	0	1	2	2.205607	41.44389
Friday	July	14	14	Afternoon	1	0	1	2	2.205118	41.44482
Sunday	January	8	20	Afternoon	0	0	0	1	2.200956	41.43713

6 rows | 6-16 of 15 columns

Vamos a analizar ahora los datos de interés de nuestro dataset. Entendemos que tenemos 3 tipologías de atributos:

- **Consecuencias de accidentes:** Mild.injuries, Serious.injuries, Victims, Vehicles.involved - Mild.injuries, Serious.injuries, Victims, Vehicles.involved.
  Tratamiento: Estos datos claramente son todos de interés, por tanto, queremos mantenerlos en el modelo.
- **Ubicación geográfica:** District.Name, Neighborhood.Name, Street, Longitude, Latitude.
  Nuestra propuesta en este caso sería prescindir de la variable Street, ya que parece que no aporta demasiado valor diferencial.
- **Ubicación temporal:** Weekday, Month, Day, Hour, Part.of.the.day.
  Nuestra propuesta sería prescindir de Hour y Part.of.the.day en favor de una nueva variable categórica Hour.Span, que pueda resultar más "explicativa". Esta variable constará de franjas horarias de 3h: 6-9h, 9-12h, 12-15h, 15-18h, 18-21h, 21-00h, 00-03h, 03-06h.
  Adicionalmente, proponemos transformar la variable Day en una variable categórica con un rango de 5 días: 1-5, 6-10, 11-15, 16-20, 21-25, 26-31.
  Por último, proponemos crear una nueva variable Weekday.Weekend a partir de Weekday, que únicamente indicará si corresponde a un día laborable o no.

En el proyecto RMD/html se pueden observar las transformaciones aplicadas sobre el dataset original.

Una vez aplicada las transformaciones planteadas, podemos observar finalmente el dataset procesado obtenido:

	Id <fctr>	District.Name <fctr>	Neighborhood.Name <fctr>	Longitude <dbl>	Latitude <dbl>	Month <fctr>	Day.Span <fctr>
1	2017S008429	Unknown	Unknown	2.125624	41.34004	October	11-15
2	2017S007316	Unknown	Unknown	2.120452	41.33943	September	1-5
3	2017S010210	Unknown	Unknown	2.167356	41.36089	December	6-10
4	2017S006364	Unknown	Unknown	2.124529	41.33767	July	21-25
5	2017S004615	Sant Martí	el Camp de l'Arpa del Clot	2.185272	41.41636	May	21-25
6	2017S007775	Sant Martí	el Camp de l'Arpa del Clot	2.183245	41.41634	September	16-20

6 rows | 1-8 of 14 columns

Hour.Span <fctr>	Weekday <fctr>	Weekday.Weekend <fctr>	Mild.injuries <int>	Serious.injuries <int>	Victims <int>	Vehicles.involved <int>
06-09h	Friday	Weekday	2	0	2	2
12-15h	Friday	Weekday	2	0	2	2
21-00h	Friday	Weekday	5	0	5	2
00-03h	Friday	Weekday	1	0	1	2
12-15h	Thursday	Weekday	1	0	1	3
12-15h	Wednesday	Weekday	1	0	1	2

6 rows | 9-15 of 14 columns

3. Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Vamos a estudiar los atributos que podrían ser problemáticos, para detectar cuáles de ellos requieren de tratamiento.

```
'data.frame':   10339 obs. of  14 variables:
 $ Id              : Factor w/ 10335 levels "2017S000001    ",..: 7990 6890 9761 5946 4210 7345 4079 10230 4747
3529 ...
 $ District.Name    : Factor w/ 11 levels "Ciutat Vella",..: 11 11 11 11 8 8 8 8 8 8 ...
 $ Neighborhood.Name: Factor w/ 74 levels "Baró de Viver",..: 69 69 69 69 12 12 12 12 12 12 ...
 $ Longitude        : num  2.13 2.12 2.17 2.12 2.19 ...
 $ Latitude         : num  41.3 41.3 41.4 41.3 41.4 ...
 $ Month            : Factor w/ 12 levels "January","February",..: 10 9 12 7 5 9 5 12 6 5 ...
 $ Day.Span         : Factor w/ 6 levels "1-5","6-10","11-15",..: 3 1 2 5 5 4 4 6 3 1 ...
 $ Hour.Span        : Factor w/ 8 levels "00-03h","03-06h",..: 3 5 8 1 5 5 8 7 6 7 ...
 $ Weekday          : Factor w/ 7 levels "Monday","Tuesday",..: 5 5 5 5 4 3 6 2 1 3 ...
 $ Weekday.Weekend  : Factor w/ 2 levels "Weekday","Weekend": 1 1 1 1 1 1 2 1 1 1 ...
 $ Mild.injuries    : int  2 2 5 1 1 1 1 2 1 1 ...
 $ Serious.injuries : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Victims          : int  2 2 5 1 1 1 1 2 1 1 ...
 $ Vehicles.involved: int  2 2 2 2 3 2 2 1 1 1 ...
```

En el proyecto RMD se puede observar cómo **no existe ningún valor NA** en ninguno de los atributos, pero en **"District.Name" y "Neighborhood.Name" existen 27 valores "Unknown"**. Por tanto, en primer lugar nos tenemos que plantear que hacer con los registros "Unknown". Podríamos optar por varias alternativas:

1. Teniendo en cuenta que la información relativa al resto de atributos puede ser interesante, se podrían **mantener estos registros** con valores no informados en el dataset.

2. Tratándose de relativamente "pocos" registros, podríamos **completar la información**: bien a mano, o bien usando algún método de imputación.

En este caso, **optamos por (2) tratar los registros usando el método KNN**, para k=1 y usando las variables longitud y latitud como variables sobre las cuales vamos a calcular la distancia. Elegimos k=1, ya que asumimos que estaremos en el mismo barrio y distrito de las coordenadas más cercanas, aunque corremos el riesgo de que se trate de unas coordenadas en la frontera de un barrio.

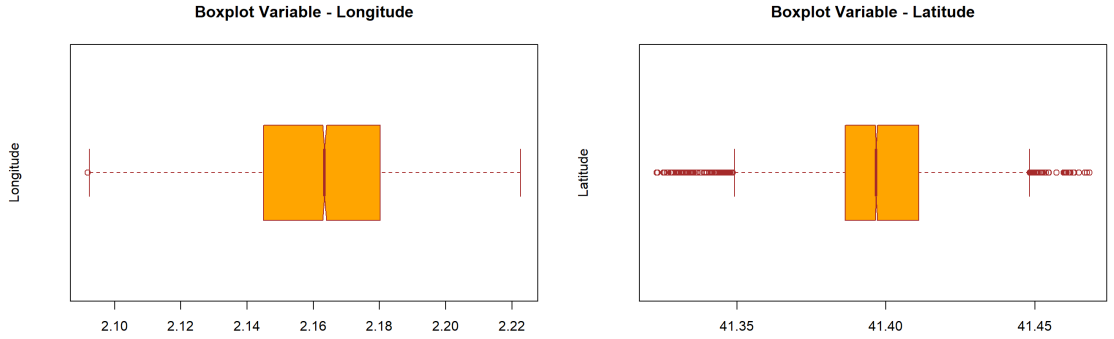
En el RMD aparece el detalle de la aplicación de esta imputación. Podemos ver a continuación como finalmente estos registros ya aparecen correctamente informados:

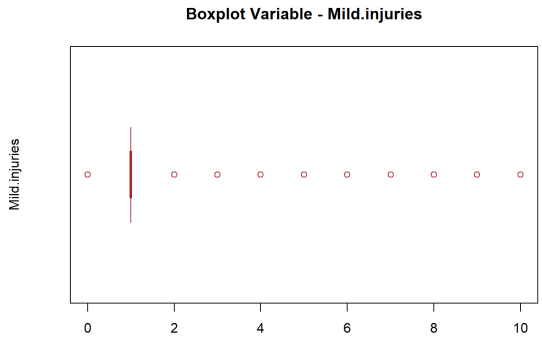
	Id <fctr>	District.Name <fctr>	Neighborhood.Name <fctr>	Longitude <dbl>	Latitude <dbl>	Month <fctr>	Day.Span <fctr>
1	2017S008429	Sants-Montjuïc	la Marina del Prat Vermell	2.125624	41.34004	October	11-15
2	2017S007316	Sants-Montjuïc	la Marina del Prat Vermell	2.120452	41.33943	September	1-5
3	2017S010210	Sants-Montjuïc	la Marina del Prat Vermell	2.167356	41.36089	December	6-10
4	2017S006364	Sants-Montjuïc	la Marina del Prat Vermell	2.124529	41.33767	July	21-25
5	2017S004615	Sant Martí	el Camp de l'Arpa del Clot	2.185272	41.41636	May	21-25
6	2017S007775	Sant Martí	el Camp de l'Arpa del Clot	2.183245	41.41634	September	16-20

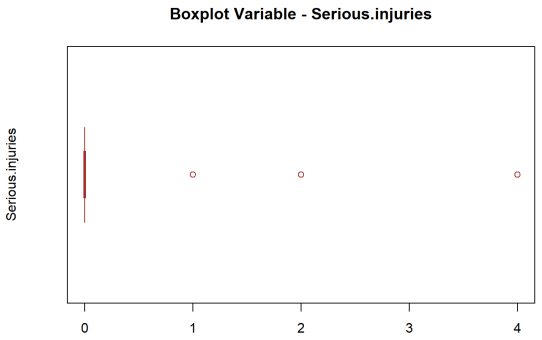
6 rows | 1-8 of 16 columns

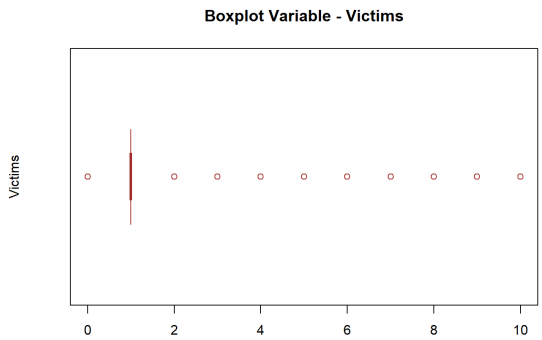
3.2 Identificación y tratamiento de valores extremos.

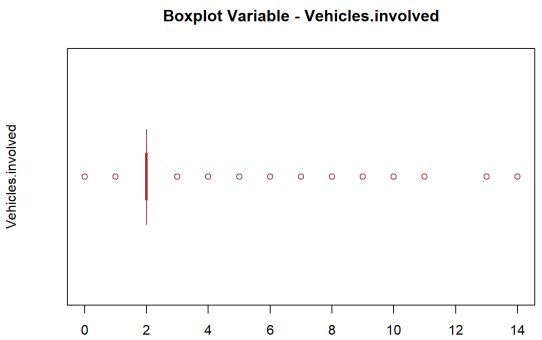
Si analizamos nuestros atributos numéricos, podemos observar que **no existen valores extremos**, y por tanto no es necesario aplicar ningún tratamiento.
Aunque el gráfico *boxplot* identifique algunos valores como outliers, dada la naturaleza de los datos y la información que representan, entendemos que no tiene sentido considerarlos como tales. Por tanto, estos datos se van a tratar como registros de datos válidos propios del dataset.











4. Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar)

Vamos a seleccionar diversas agrupaciones de datos que a priori pueden ser de interés en posteriores apartados. Es un planteamiento inicial, de forma que es posible que no acabemos usando todos estos subconjuntos en apartados posteriores.

Agrupación por día festivo o laborable

```
accidentes.laborable <- df_proc[df_proc$Weekday.Weekend == "Weekday",]
accidentes.festivo <- df_proc[df_proc$Weekday.Weekend == "Weekend",]
```

Agrupación por dia laborable "normal" o viernes

```
accidentes.lun_vie <- df_proc[df_proc$Weekday %in% c("Monday", "Tuesday", "Wednesday",
"Thursday"),]
accidentes.viernes <- df_proc[df_proc$Weekday == "Friday",]
```

Agrupación por meses de verano o resto año

```
`%notin%` <- Negate(`%in%`)
accidentes.verano <- df_proc[df_proc$Month %in% c("July", "August"),]
accidentes.no_verano <- df_proc[df_proc$Month %notin% (c("July", "August")),]
```

Agrupación por primeros o últimos dias del mes

```
accidentes.ppioMes <- df_proc[df_proc$Day.Span == "1-5",]
accidentes.finMes <- df_proc[df_proc$Day.Span == "26-31",]
```

Agrupación por franjas horarias particulares

```
accidentes.madrugada <- df_proc[df_proc$Hour.Span %in% c("00-03h", "03-06h"),]
accidentes.primeraHora <- df_proc[df_proc$Hour.Span == "06-09h",]
accidentes.afterWork <- df_proc[df_proc$Hour.Span == "18-21h",]
accidentes.noche <- df_proc[df_proc$Hour.Span == "21-00h",]
accidentes.restoHoras <- df_proc[df_proc$Hour.Span %in% c("09-12h", "12-15h", "15-18h"),]
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Vamos a aplicar el **método Kolmogorov-Smirnov** para comprobar la normalidad del conjunto de datos. En particular, nos vamos a fijar en nuestras variables numéricas *target*: heridos leves, heridos graves, víctimas y vehículos implicados.

```
ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  df_proc$Mild.injuries
D = 0.38844, p-value < 2.2e-16
alternative hypothesis: two-sided

ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  df_proc$Serious.injuries
D = 0.53501, p-value < 2.2e-16
alternative hypothesis: two-sided

ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  df_proc$Victims
D = 0.39753, p-value < 2.2e-16
alternative hypothesis: two-sided

ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  df_proc$Vehicles.involved
D = 0.35506, p-value < 2.2e-16
alternative hypothesis: two-sided
```

En este caso, observamos que el **p-valor es menor que el nivel de significancia**, por tanto, **no podemos asumir la normalidad de la distribución** de las variables.
Por otra parte, por **teorema del límite central** sabemos que, teniendo N suficientemente grande -como es el caso-, se puede asumir que el dataset es aproximadamente normal. Esto es una cuestión a considerar en la posible aplicación de t-test o métodos similares.

Podemos hacer una verificación de los resultados usando el test de **Anderson-Darling**, que nos dice qué variables son normales.
Los resultados son:

```
Test de Anderson-Darling: lista de variables que no siguen una distribución normal:
Longitude, Latitude, Mild.injuries, Serious.injuries,
Victims, Vehicles.involved
```

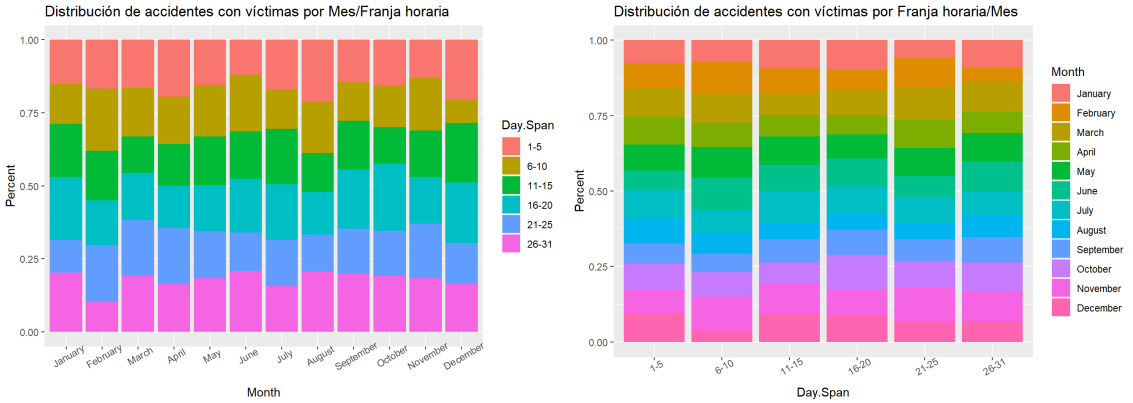
4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

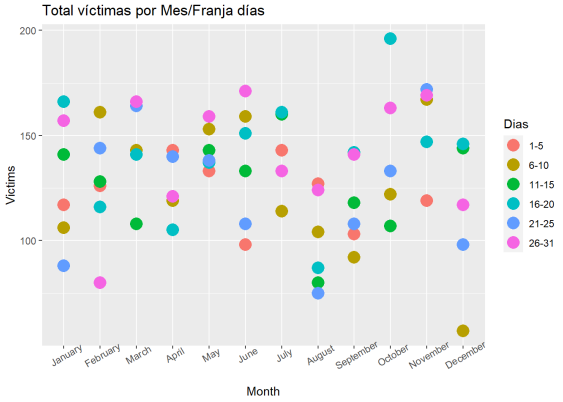
Identificar puntos negros en la ciudad

El tratamiento a esta cuestión se aborda en el apartado (5).

Identificar relación entre Meses/Días con mayor siniestralidad

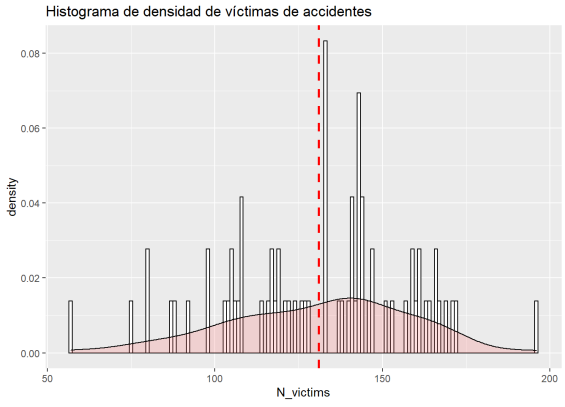
Para tratar de responder a esta cuestión, vamos a basarnos en los atributos Month y Day.Span; y observar la relación de estos con Victims.





Por lo que se observa en los anteriores gráficos, **no existe una particular correlación o una tendencia** destacable que relacione el mes/franja de días, con la siniestralidad de los accidentes registrados.

Mediante el **test Chi-Square**, vamos a analizar si estadísticamente confirma lo que observamos en los gráficos.
En primer lugar, vamos a convertir la variable Victims en categórica, para hacerlo estudiamos su distribución:



Parece que esta se puede estructurar en *buckets* de 40, de la siguiente forma:

- De 50-89 – Baja Siniestralidad
- De 90-129 – Media Siniestralidad
- De 130-169 – Alta Siniestralidad
- De 170-2109 – Muy Alta Siniestralidad

Una vez codificada de forma categórica nuestra variable target, aplicamos el test Chi-Square:

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  df_sin_chisq$Month and df_sin_chisq$Victim.Span
X-squared = 44.488, df = 33, p-value = 0.08739

Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  df_sin_chisq$Day.Span and df_sin_chisq$Victim.Span
X-squared = 9.4021, df = 15, p-value = 0.8556
```

Al obtener un **p-valor superior al nivel de significación**, no podemos rechazar la hipótesis nula. De manera que **no se puede afirmar que las variables sean dependientes**.
Por tanto, como indicaban los gráficos, parece que no existe una relación entre la siniestralidad y la distribución en cuanto a Mes/Franja de días. Únicamente a comentar que, la más cercana al valor de significación es el atributo Month, por tanto es la más "cercana" a ser significativa.

Identificar si existe mayor siniestralidad a principios o finales de mes

Para abordar esta cuestión, vamos a aplicar un contraste de hipótesis entre los frames temporales de principio y fin de mes. Para esto vamos a aplicar el **t.test**, que como comentamos anteriormente, podemos aplicar en base al teorema del límite central.

Vamos a usar las agrupaciones definidas en (4.1): accidentes.ppioMes y accidentes.finMes. El resultado obtenido de la aplicación de t.test para nuestras variables *target* es:

```
data:  accidentes.ppioMes$Mild.injuries and accidentes.finMes$Mild.injuries
t = -0.83572, df = 3524.5, p-value = 0.4034

data:  accidentes.ppioMes$Serious.injuries and accidentes.finMes$Serious.injuries
t = -0.061248, df = 3432.2, p-value = 0.9512

data:  accidentes.ppioMes$Victims and accidentes.finMes$Victims
t = -0.82604, df = 3507.3, p-value = 0.4088

data:  accidentes.ppioMes$Vehicles.involved and accidentes.finMes$Vehicles.involved
t = 0.2687, df = 3501.4, p-value = 0.7882
```

Observamos que en ningún caso obtenemos un p-valor inferior del nivel de significación, por tanto, no podemos rechazar la hipótesis nula. Así pues, **el hecho de que sea principio o final de mes no tiene significación estadística** en cuanto a la tipología de heridos, las víctimas o los vehículos implicados.

Identificar si existe correlación entre heridos graves en función de horario nocturno o diurno

En este caso, de nuevo vamos a usar el contraste de hipótesis con el t.test para analizar la cuestión planteada.

Para esto, vamos a construir los siguientes subgrupos a partir de nuestro dataset:

```
accidentes.noche <- df_proc[df_proc$Hour.Span %in% c("06-09h", "09-12h", "12-15h", "15-18h", "18-21h"),]
accidentes.dia <- df_proc[df_proc$Hour.Span %in% c("21-00h", "00-03h", "03-06h"),]
```

Y en base a estos, aplicamos el t.test sobre las variables Serious.Injuries y Victims. Obtenemos el siguiente resultado:

```
        Welch Two Sample t-test

data:  accidentes.noche$Serious.injuries and accidentes.dia$Serious.injuries
t = -2.3696, df = 2116.6, p-value = 0.01789
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.021913837 -0.002067362
sample estimates:
 mean of x  mean of y
0.02136259 0.03335319


        Welch Two Sample t-test

data:  accidentes.noche$Victims and accidentes.dia$Victims
t = -0.21828, df = 2100.5, p-value = 0.8272
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05009977  0.04006411
sample estimates:
mean of x mean of y
 1.177829  1.182847
```

Observamos que, para heridos graves, el **p-valor es inferior al nivel de significación**, por tanto podemos rechazar la hipótesis nula, y afirmar que **la hora del día**, respecto a los heridos graves de accidentes, **es estadísticamente significativa**.

Por contra, para el caso de accidentes con víctimas, no podemos rechazar hipótesis nula. Por tanto, no existe relación de la hora del día con las víctimas.

Identificar si hay la misma frecuencia de accidentes múltiples en fin de semana que entre semana

Definimos accidente múltiple cuando hay dos o más vehículos implicados.

Para esto, vamos a construir dos datasets con la frecuencia de accidentes múltiples por día a partir de nuestro dataset inicial:

```
df_multi_acc <- df_raw[df_raw$Vehicles.involved>1,][c("Id","Weekday","Month","Day")]
df_multi_acc_freq <- count(df_multi_acc,vars=c("Weekday","Month","Day"))
```

Como vamos a comparar la frecuencia de accidentes diarios entre semana con la del fin de semana, debemos primero ver que podemos aplicar el Teorema central del límite para asumir normalidad en las muestras. Como ambas muestras tienen más de 100 observaciones, podemos asumir normalidad.

Ahora vamos a hacer un var.test para ver si podemos o no asumir igualdad de varianzas:

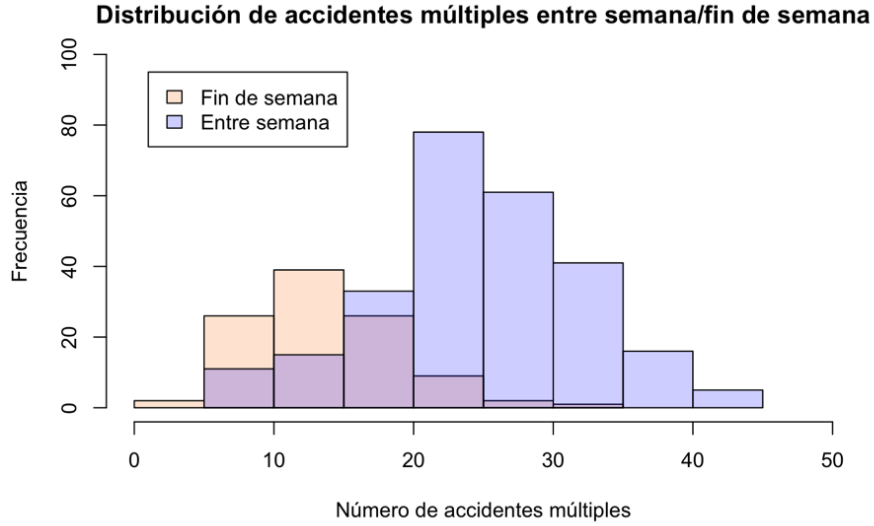
```
        F test to compare two variances

data:  acc.multi.finde$freq and acc.multi.semana$freq
F = 0.5232, num df = 104, denom df = 259, p-value = 0.0001994
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3830094 0.7309997
sample estimates:
ratio of variances
         0.5231969
```

Como tenemos p-value inferior a 0.05, no podemos asumir igualdad de varianzas.

En este caso, de nuevo vamos a usar el contraste de hipótesis con el t.test sobre la media para dos muestras con varianza desconocida para analizar si los accidentes múltiples se producen con igual frecuencia entre semana que en fin de semana:

```
        Welch Two Sample t-test

data:  acc.multi.finde$freq and acc.multi.semana$freq
t = -16.025, df = 263.48, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.17019  -9.50673
sample estimates:
mean of x mean of y
 14.40000  25.23846
```

Vemos que no, que entre semana se dan con mucha mayor frecuencia accidentes múltiples ya que el p-value es muy inferior al nivel de significancia.



Identificar si hay la misma frecuencia de accidentes en periodo vacacional que fuera de él

Definimos periodo vacacional a los meses de julio y agosto.
Para esto, vamos a construir dos datasets con la frecuencia de accidentes por día a partir de nuestro dataset inicial:

```
df_ accidentes.freq <- count(df_raw,vars=c("Month","Day"))

accidentes.verano.freq <- accidentes.freq[accidentes.freq$Month %in% c("July","August"),]
accidentes.no_verano.freq <- accidentes.freq[!(accidentes.freq$Month %in%
c("July","August")),]
```

Como vamos a comparar la frecuencia de accidentes diarios dentro y fuera del periodo vacacional, debemos primero ver que podemos aplicar el Teorema central del límite para asumir normalidad en las muestras. Como ambas muestras tienen más de 50 observaciones, podemos asumir normalidad.

Ahora vamos a hacer un var.test para ver si podemos o no asumir igualdad de varianzas:

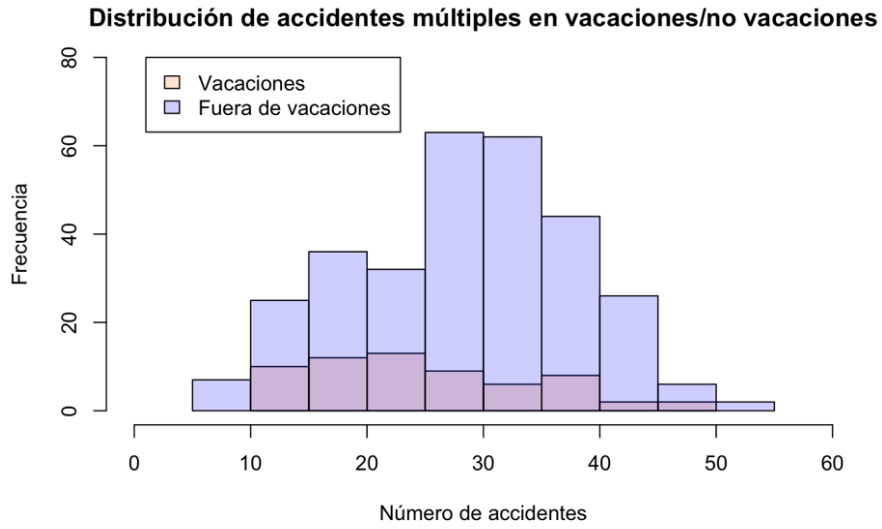
```
        F test to compare two variances

data:  accidentes.verano.freq$freq and accidentes.no_verano.freq$freq
F = 1.016, num df = 61, denom df = 302, p-value = 0.9014
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7044089 1.5427657
sample estimates:
ratio of variances
          1.016001
```

En este caso, el p_value y el intervalo de confianza nos dicen que, a diferencia del caso anterior, ahora podemos asumir igualdad de varianzas.

En este caso, de nuevo vamos a usar el contraste de hipótesis con el t.test sobre la media para dos muestras, pero con igualdad de varianzas, para analizar si los accidentes múltiples se producen con igual frecuencia en periodo vacacional o fuera de él:

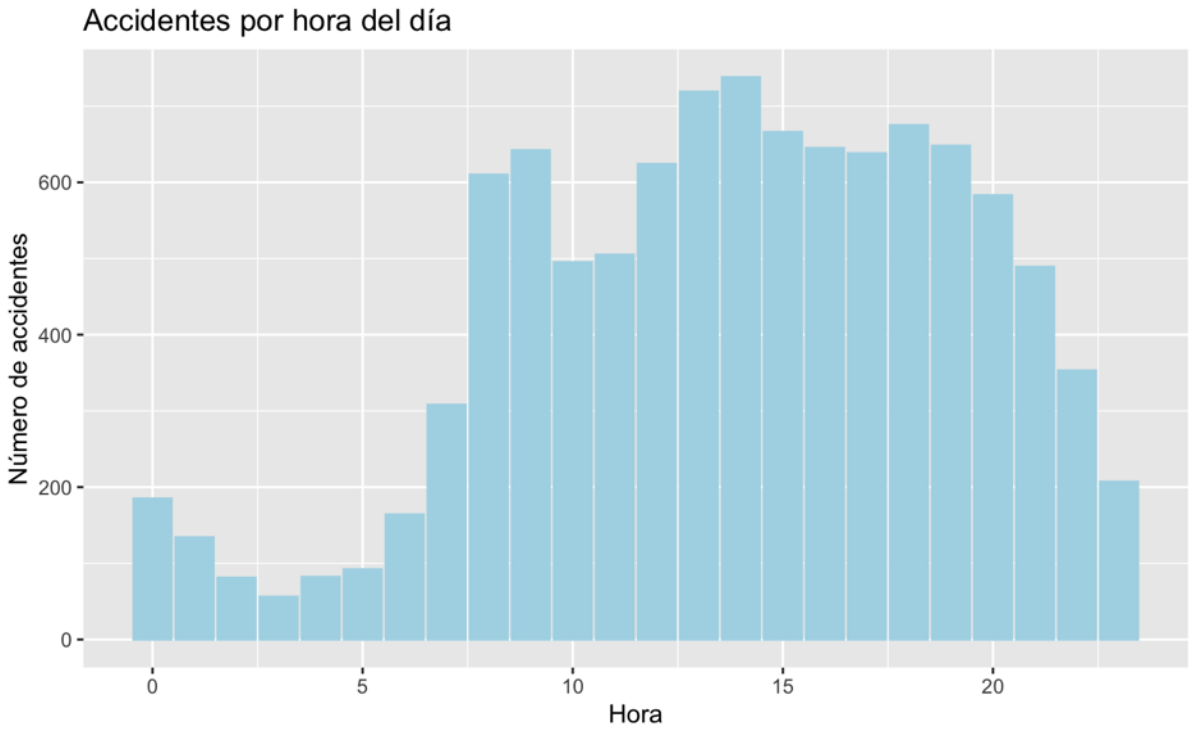
```
       Two Sample t-test

data:  accidentes.verano.freq$freq and accidentes.no_verano.freq$freq
t = -2.6917, df = 363, p-value = 0.007438
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.2612752 -0.9747516
sample estimates:
mean of x mean of y
 25.32258  28.94059
```

El test nos muestra que la frecuencia de accidentes en verano es inferior a la del resto del año.

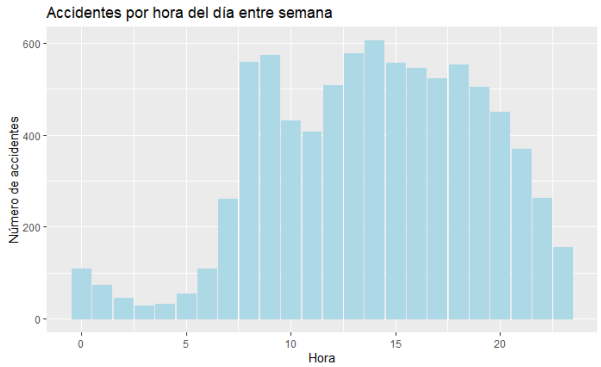


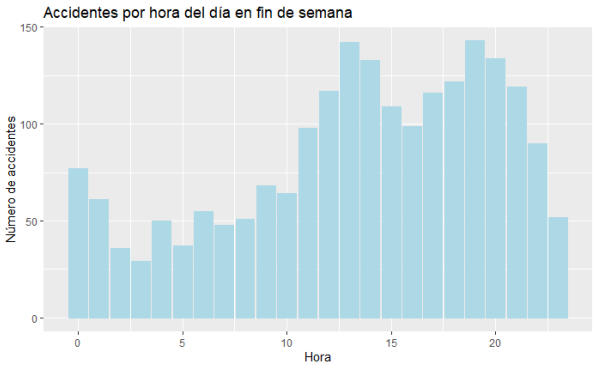
Identificar si existe correlación entre la hora y el número de accidentes

Al obtener un histograma de accidentes por horas, se observa que las horas con mayor siniestralidad son los desplazamientos al trabajo por la mañana (las 8 y las 9h), los desplazamientos a mediodía (13 y 14h) y, de manera más generalizada, la vuelta del trabajo a casa por la tarde, hasta prácticamente las 21h.

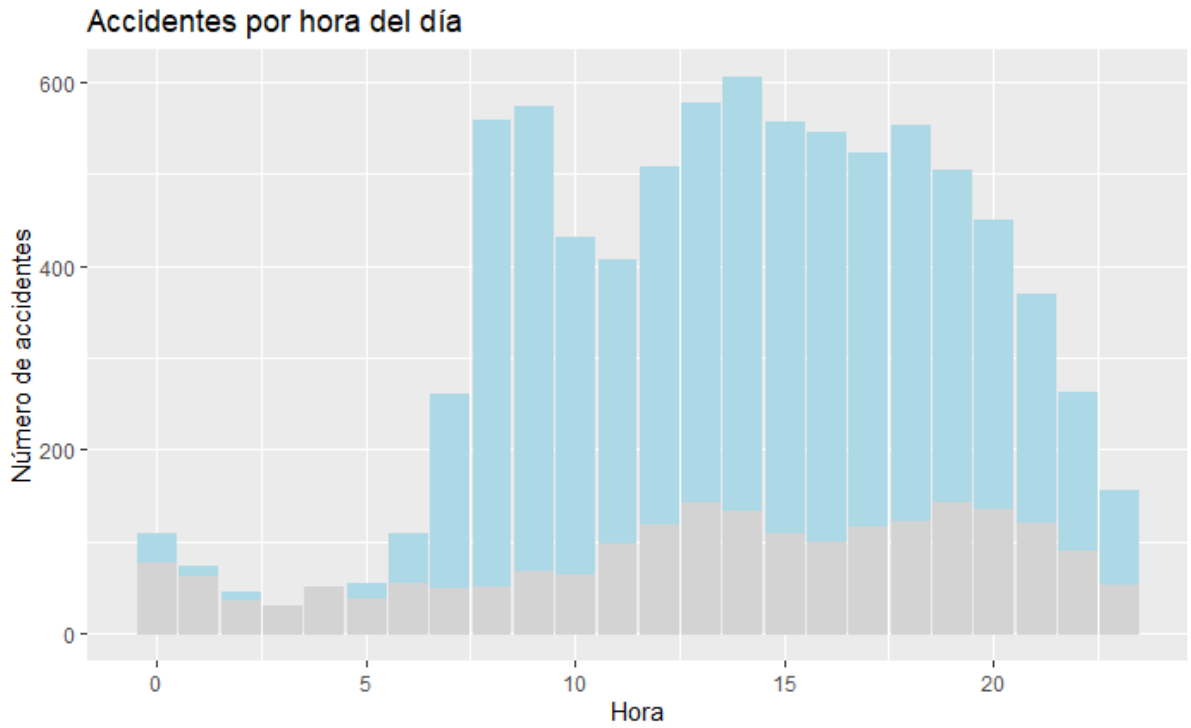


Separando el dataset en accidentes entre semana y fines de semana, vemos que la distribución cambia. En fin de semana, como es previsible, desaparece el pico de accidentes matinales (8 a 10h) y se producen picos en las horas de comer y cenar.



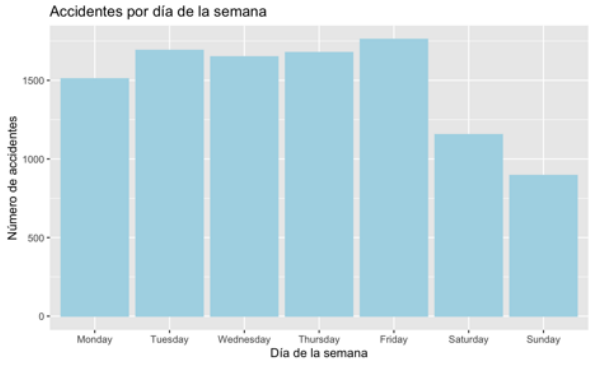


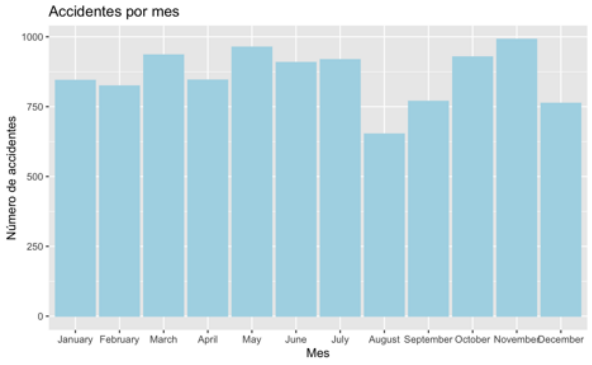
Superponiendo los dos gráficos, vemos que la incidencia de accidentes baja muchísimo en fin de semana:

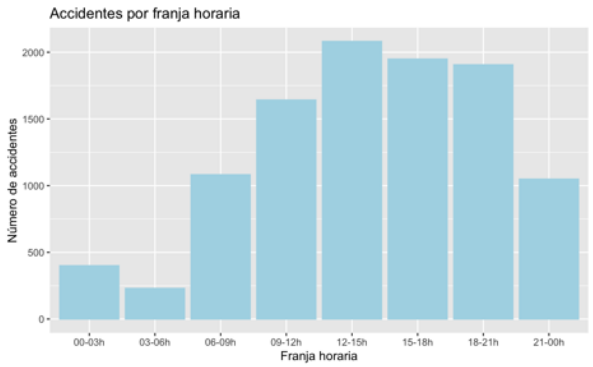


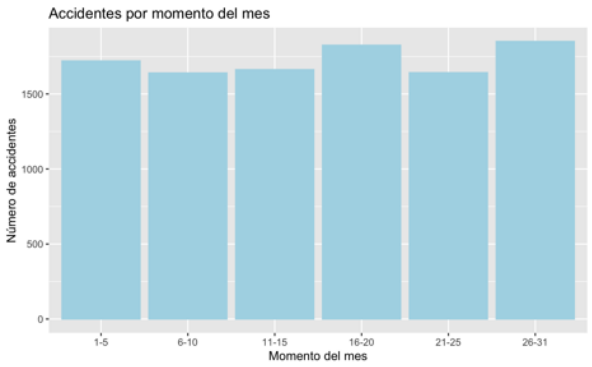
5. Representación de los datos a partir de tablas y gráficas

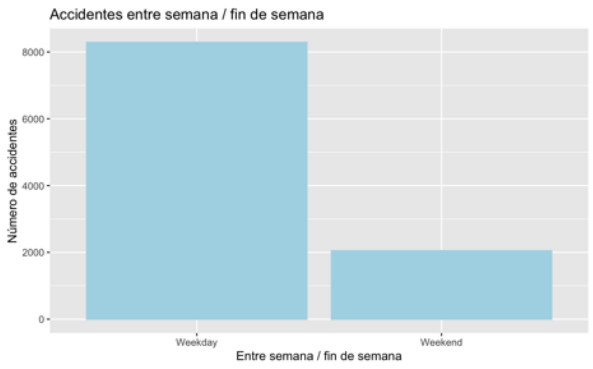
5.1 Análisis preliminar de los datos mediante gráficas/histogramas con sus atributos

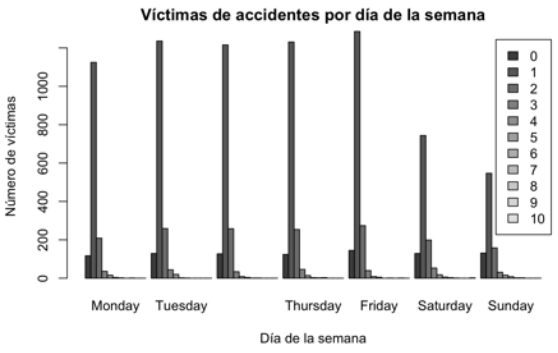


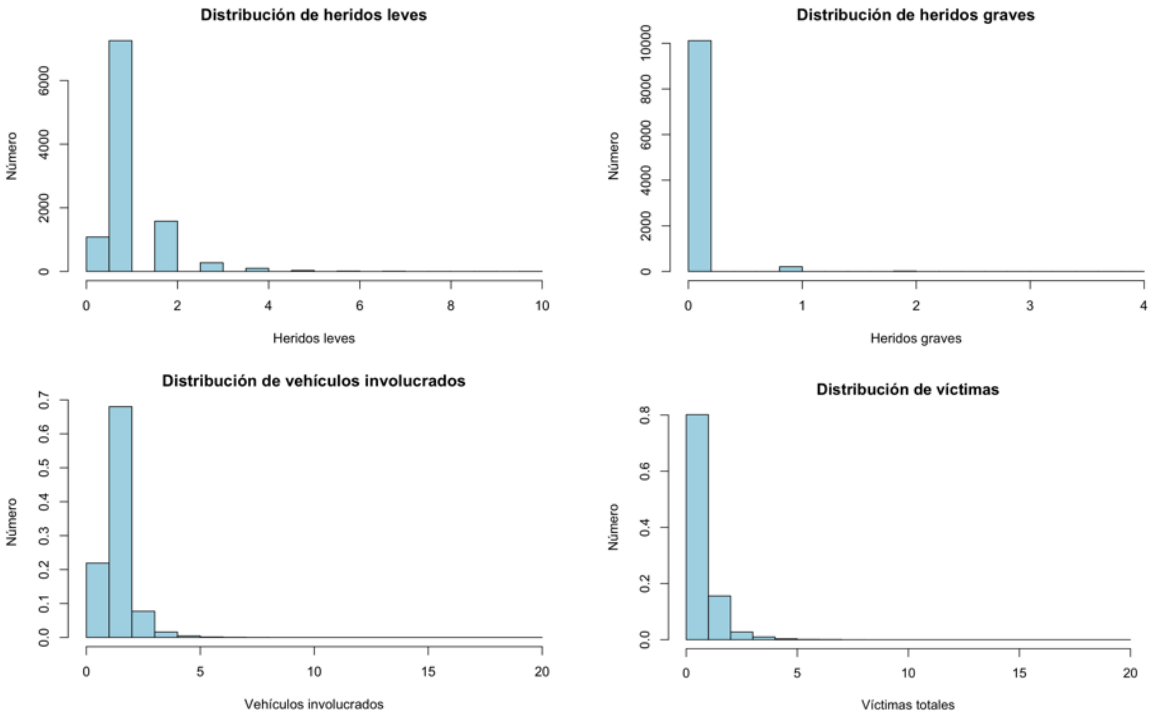










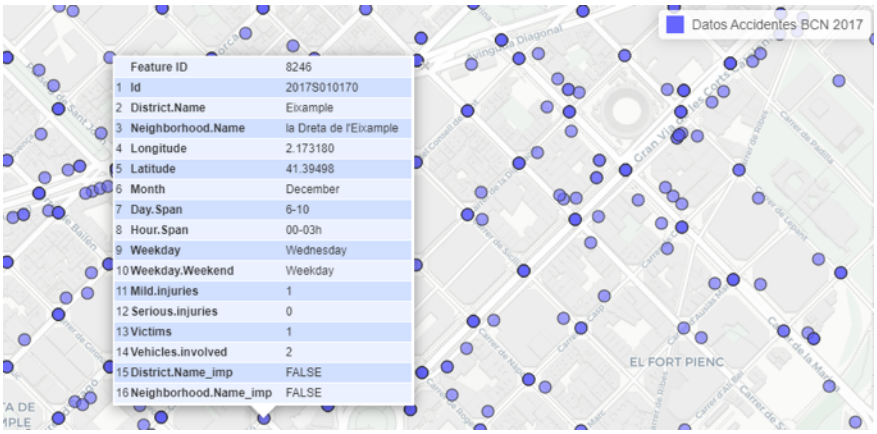


5.2 Visualización de puntos negros de la ciudad

El dataset contiene en total **10.339 registros** de accidentes con sus distintas ubicaciones geolocalizadas con longitud y latitud.

Hemos usado la *library* **mapview** y **mapshot** de R por tal de realizar lo siguiente:

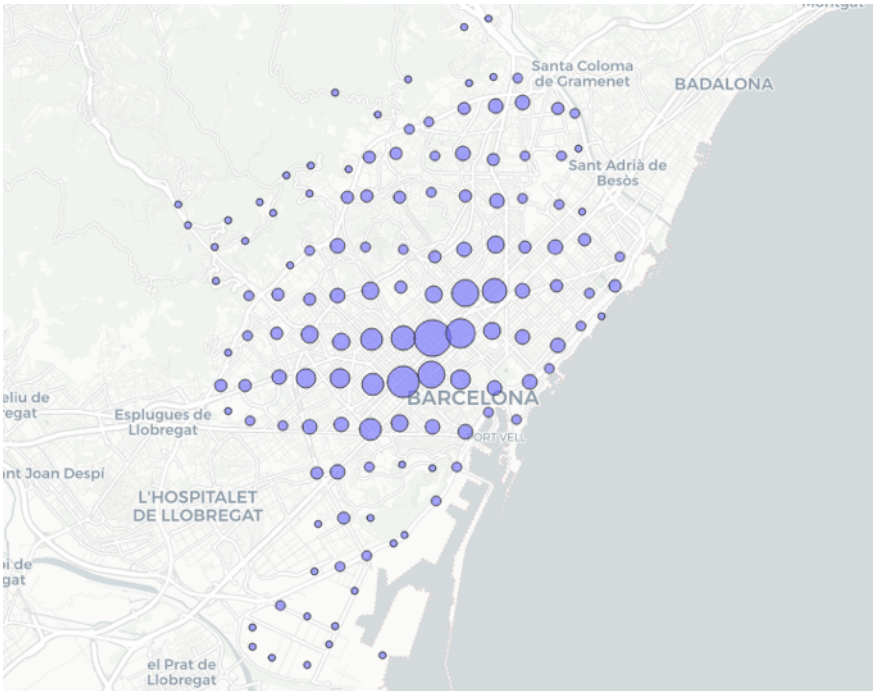
- Generar un HTML con un **mapa interactivo del dataset completo**



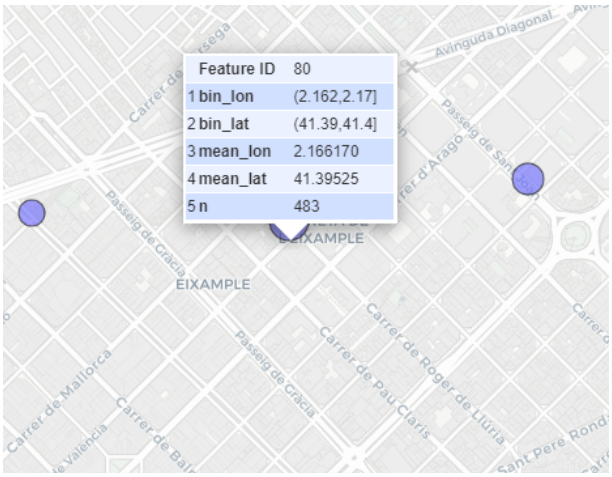
- Generar un HTML con un **mapa interactivo agrupado por una grid de 15x15**
  El objetivo de esto era disponer de un mapa que en una vista general permitiera hacer una observación de la distribución de accidentes.

- Generar una imagen PNG con una foto fija del **mapa agrupado por una grid de 15x15**
  El objetivo de esto es poder incluir una representación estática en esta memoria, o en otra documentación.

El resultado es este:



Esta visualización nos permite observar como el **punto con mayor concentración de accidentes** se encuentra en el **centro de Barcelona** (Pg. Gracia/Pg. Sant Joan entre Diagonal/Gran vía), con un total de 483 accidentes acumulados a lo largo del 2017.



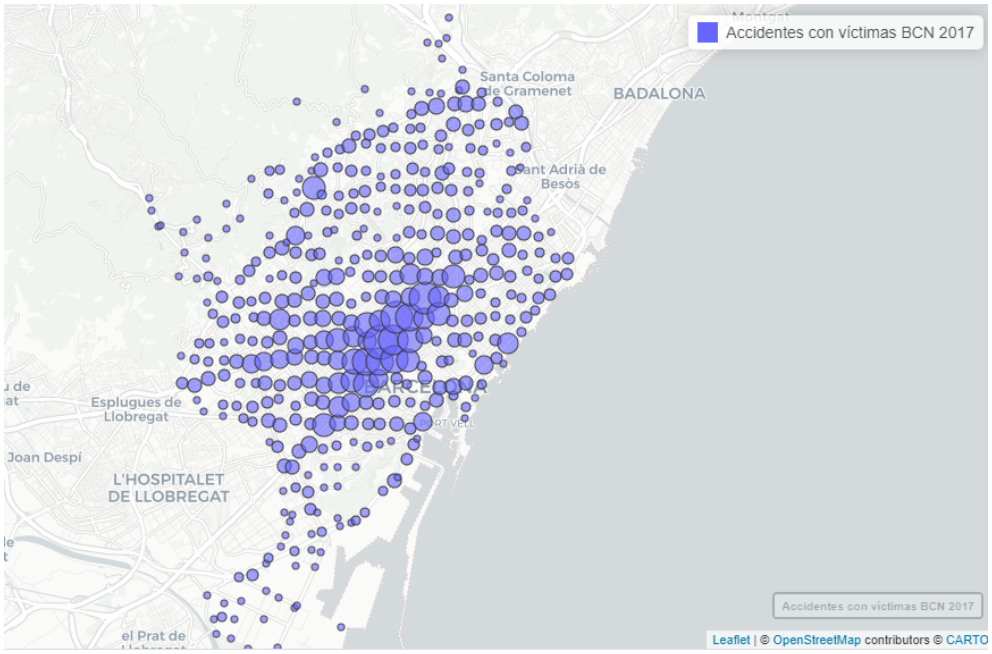
De la misma forma, la visualización nos permite observar otras zonas con una distribución importante:

1. La zona centro en general de Barcelona
2. Las Rondas de la ciudad (especialmente Ronda de Dalt, y Ronda Litoral)

Adicionalmente, como esta visualización representa una información interesante a nivel de la localización de los accidentes, hemos extendido el estudio a otros casos:

Visualización de ubicaciones de accidentes con víctimas

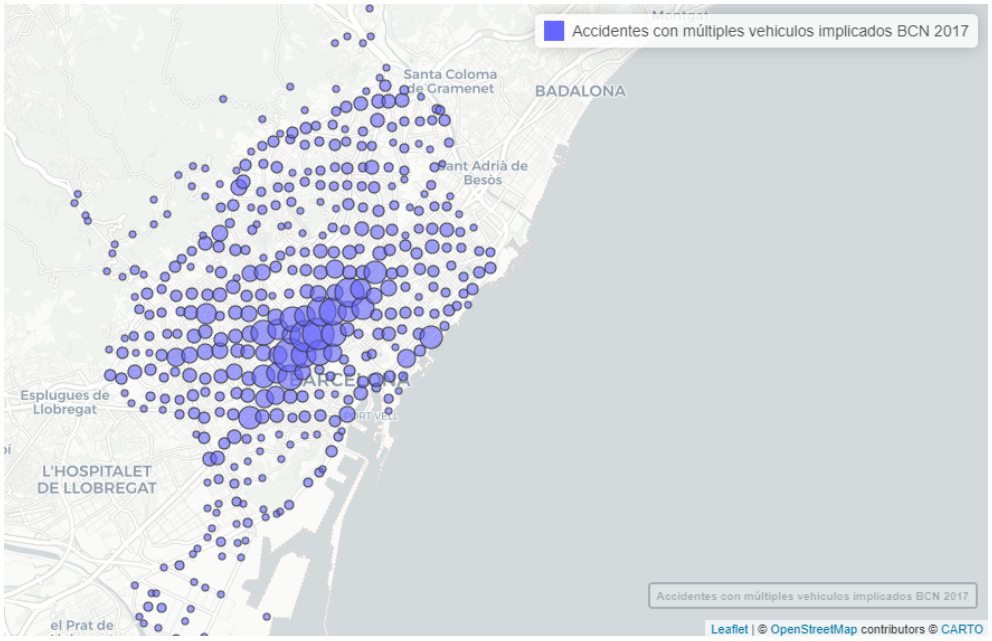
En este caso, al tratarse de menor número de casos, hemos decidido hacer una grid más amplia (de 30x30), obteniendo el siguiente resultado:



Se observa una **tendencia muy similar** a la del dataset general, con tres ejes pronunciados: zona centro de Barcelona, y de nuevo la Ronda de Dalt y Ronda Literal.

Visualización de ubicaciones de accidentes con múltiples vehículos implicados

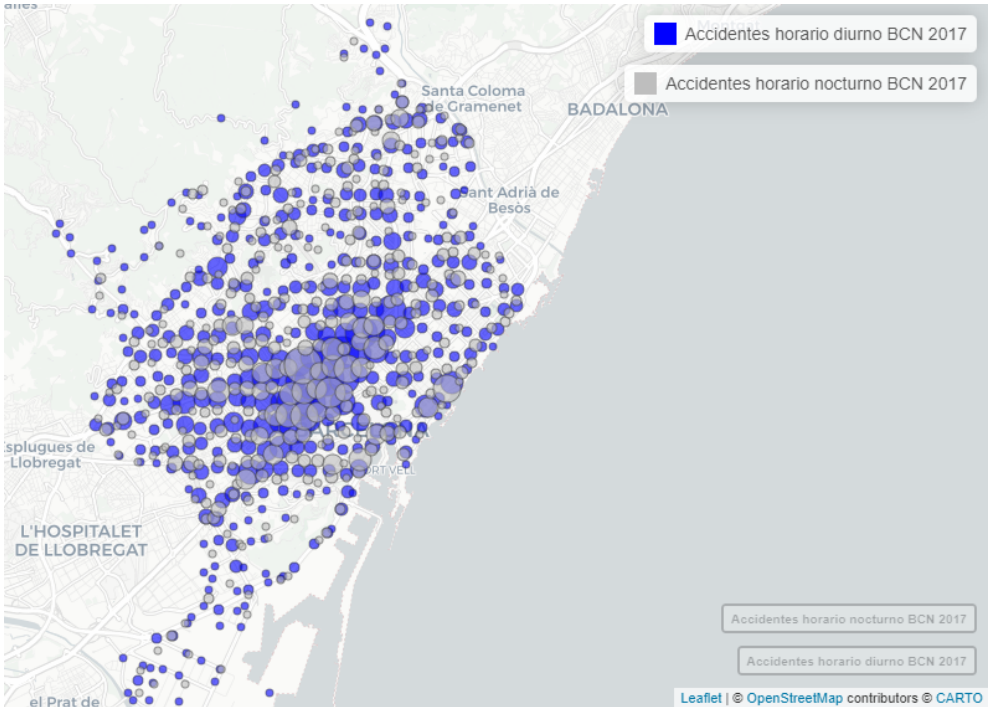
En este caso, de nuevo usamos una grid de 30x30, y representamos únicamente los accidentes con más de un vehículo implicado. Obtenemos el siguiente resultado:



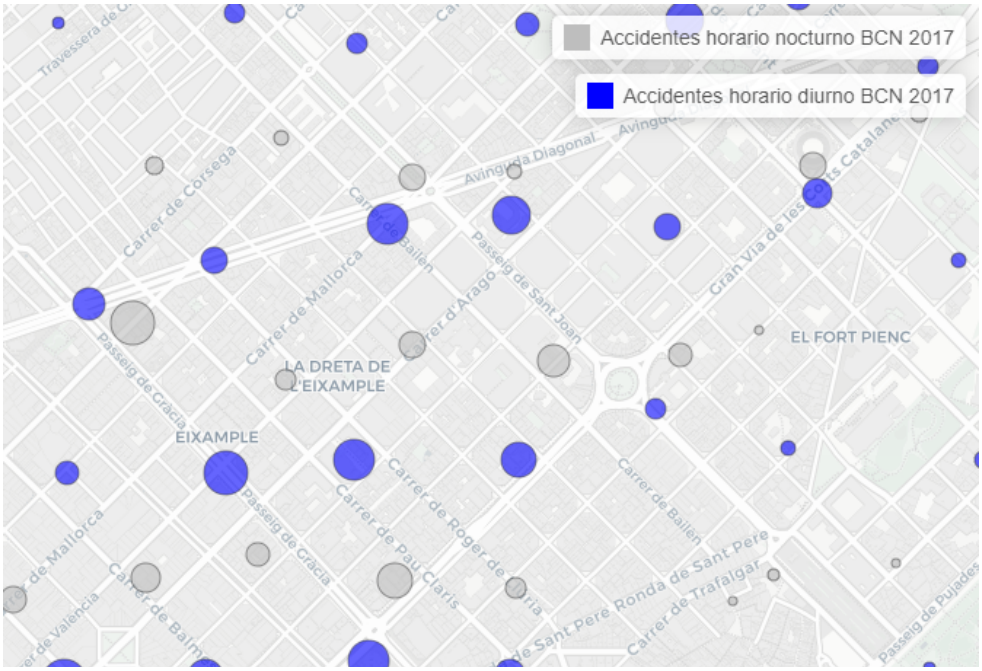
De nuevo, observamos una **tendencia muy similar** a los casos anteriores, por tanto en base a este dataset no hay información destacable a extrapolar en cuanto casos de accidentes múltiples.

<u>Visualización de ubicaciones de accidentes en horario diurno/nocturno</u>

En este caso, con una grid de 30x30, representamos los accidentes en horario diurno contrastados con los de horario nocturno. El resultado obtenido es el siguiente:



En este caso, si hacemos zoom en la zona centro de Barcelona, sí observamos ligeras diferencias a tener en cuenta:



Esta zona, como habíamos visto antes, era la que tenia mayor concentración de accidentes. Observamos en este caso que existe una sensible diferencia en los casos entre día/noche, probablemente dado por que es una zona de oficinas/trabajo, **desplazando la concentración de accidentes hacia Paseo de Gracia en horario nocturno**.

También se observan en algunos otros puntos, por ejemplo, la Ronda Litoral en la zona de Vil·la Olímpica o Port Vell, con un cierto incremento en la siniestralidad en horario nocturno. Esto de nuevo nos encaja, dado que son zonas claramente con movimiento asociado al ocio nocturno.

6. Resolución del problema

1. Destacar que estamos ante un dataset de "**datos reales**", esto hace que los resultados no muestren información o tendencias claras o muy pronunciadas. Esto contrasta con otros dataset "clásicos" de aprendizaje como puede ser el wine, iris, etc. que quizás si permiten extraer conclusiones más directas y claras.

2. En cuanto a **puntos negros**, se puede observar una importante **concentración en determinados puntos de la ciudad**. De manera que quizás valdría la pena realizar un estudio a fondo para determinadas áreas, y realizar un análisis a nivel de organización de tráfico, para ver si es posible aportar una cierta mejora.

3. En los análisis planteados, hemos encontrado **relación significativa entre la hora del día y accidentes con heridos graves**. Esto apuntaría a la mayor peligrosidad de esta franja horaria, y de nuevo combinado con determinada área de influencia, podrá dar lugar a tomar medidas al respecto: mayores controles policiales, ubicación de radares de velocidad, etc.

4. Analizando la cantidad de accidentes que se producen en la ciudad entre semana y fin de semana, se observa que su distribución es diferente en ambos periodos de tiempo, siendo mucho **más frecuentes los accidentes durante la semana**. Este resultado es el esperado teniendo en cuenta la cantidad de desplazamientos que se producen para ir al lugar de trabajo o estudio. La métrica interesante sería ver número de accidentes con respecto a número de desplazamientos, pero con los datos presentes en el dataset, no es posible.

5. Como en el punto anterior, la lógica se impone de nuevo y nos muestra que la **distribución de accidentes en periodo vacacional es menor** que fuera de él, también claramente ligada a los desplazamientos al trabajo o escuela/universidad. Igualmente, disponer de cifras de desplazamientos serviría para contextualizar este resultado.

6. La distribución de los accidentes según la hora del día también muestra diferencias en cuanto a la distribución, siendo **más comunes los accidentes en horarios de desplazamientos** (mañana, mediodía y tarde), si bien los accidentes a primera hora están más concentrados y por la tarde más distribuidos.
Haciendo el ejercicio de separar accidentes entre semana y en fin de semana, vemos que la distribución es ligeramente diferente, sobre todo porque en fines de semana desaparecen los accidentes de la mañana pasando a concentrarse a mediodía y por la noche, coincidiendo con los desplazamientos para encuentros sociales (comida, cena) y ocio nocturno.

Por último, comentar **siguientes pasos** que podrían enriquecer el estudio de este dataset:
Para acabar de extraer conclusiones más potentes de este dataset, nos haría falta una base general sobre la que establecer los datos: la **distribución del volumen de tráfico** en Barcelona, tanto por sectores como por horas/meses. Esto nos permitiría estudiar la información con mayor contexto.
Adicionalmente, relacionado con los accidentes en horario nocturno, entendemos que poder disponer de información en cuanto a **controles de alcoholemia** registrados por ubicación y fecha nos permitiría añadir una nueva dimensión, y aportar mayor valor a este estudio

Participación

Contribuciones	Firma
Investigación previa	Víctor María Cardoner Álvarez José Oriol Bielsa Nogaledo
Redacción de las respuestas	Víctor María Cardoner Álvarez José Oriol Bielsa Nogaledo
Desarrollo del código	Víctor María Cardoner Álvarez José Oriol Bielsa Nogaledo