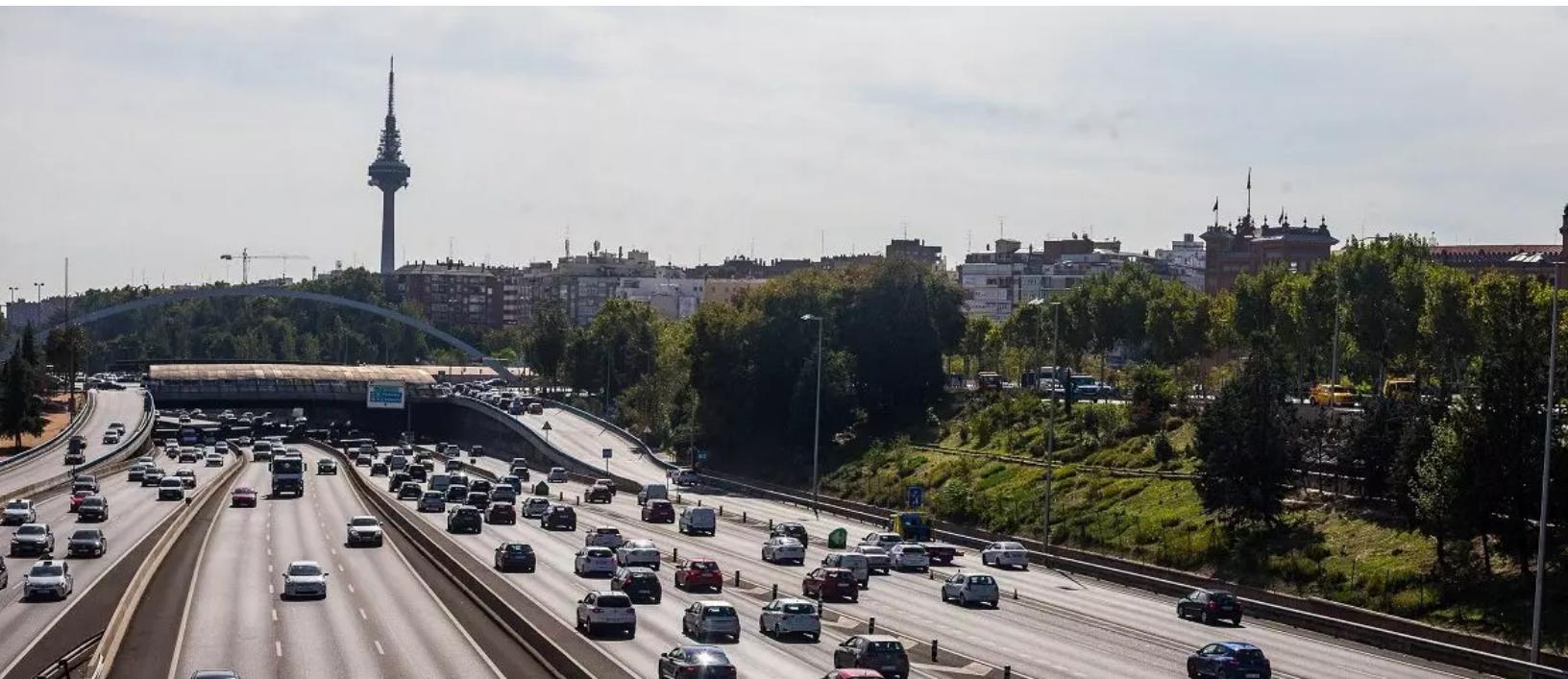


## **PROYECTO II**

# **Estudio de la siniestralidad de tráfico en Madrid (2019-2022)**



**Ciencia de Datos — Curso 2022/2023**

Coral Montes, Adrián Rico, Juan Tomás, Marc Vicedo, Tingting Wu

---

# Índice

- Alcance del proyecto
  - Objetivos planteados
  - Utilidad del estudio
  - Obtención de los datos
- Desarrollo
  - **Objetivo 1:** Estudio por criterios temporales
  - **Objetivo 2:** Estudio por ubicación
  - **Objetivo 3:** Influencia de la meteorología
  - **Objetivo 4:** Contraste turismo-moto
  - **Objetivo 5:** Influencia del COVID-19
- Visión de futuro
- Anexos
  - **Anexo 1:** Captura de los datos
  - **Anexo 2:** Integración de los datos
  - **Anexo 3:** Figuras gráficas
  - **Anexo 4:** Ficheros R Markdown

# Alcance y configuración

Con la elaboración de esta memoria, buscamos presentar un análisis profundo de los accidentes de tráfico que se producen en el área metropolitana de Madrid. Para ello, hemos estudiado todos los siniestros viales ocurridos en dicha zona durante los cuatro últimos años, para así contar con un conjunto amplio de datos de diversos tipos con el que trabajar.

---

## Objetivos

Establecemos los siguientes objetivos para desarrollar nuestro estudio:

1. Caracterizar los accidentes en Madrid en función de **criterios temporales**, observando la variación de frecuencia de los accidentes en función del calendario (días festivos, vacaciones, partidos de fútbol...).
2. Observar la variación de frecuencia de accidentes en función de la **ubicación** de los mismos en los distintos distritos de Madrid. A partir del volumen total de accidentes de tráfico por distritos, estudiar los distritos con mayor volumen de accidentes, analizando en qué zonas se concentran y posibles hechos que han causado dicho volumen.
3. Observar si la **meteorología** influye en nuestra forma de conducir: analizar si la siniestralidad de tráfico es mayor en los días con mal tiempo.
4. Contrastar el desenlace de los accidentes de **turismos y motos**; encontrar patrones en clases de vehículo (no abundantes), como geolocalización. Frecuencia de los accidentes sobre los puntos negros: tiempo(días de la semana), ver donde se dan más accidentes, localización de los accidentes...
5. Plasmar la **influencia del COVID-19** en la movilidad y en la accidentalidad, comparando los datos del año 2020 con los de los años 2019, 2021 y 2022.

## Utilidad del estudio

Hemos distinguido cuatro grupos poblacionales a los que consideramos que podría serles útil nuestro proyecto:

- A. El **Ayuntamiento** de la ciudad podría estar interesado en nuestro análisis, ya que este puede ayudar a mejorar la movilidad en el área metropolitana de Madrid, al identificar las zonas en las que aparecen más incidencias de tráfico y a qué se deben estas.
  - B. También podría ser útil para **compañías de seguros**, ya que, si identificamos en qué franjas horarias y qué tipos de conductores tienen más accidentes, podrían ajustar los precios a esos tipos de clientes.
  - C. Con nuestro análisis, la **Dirección General de Tráfico** podría controlar mejor las zonas más conflictivas en lo que a siniestralidad se refiere, e intentar prevenir accidentes reforzando la seguridad en dichos puntos.
  - D. Finalmente, el estudio puede ser útil para los propios **conductores y peatones**, de modo que puedan evitar (o al menos desplazarse con más precaución) las vías donde se producen mayor número de accidentes al darse ciertas características, tales como la meteorología, la afluencia de tráfico o el día de la semana.
- 

## Obtención de los datos

Nuestros datos son de titularidad pública (es decir, datos abiertos), y han sido obtenidos de tres organismos públicos españoles: el Ayuntamiento de Madrid, la Dirección General de Tráfico y la Agencia Estatal de Meteorología. La adquisición de estos datos ha sido totalmente legal, y al no incluirse atributos sensibles ni ningún tipo de identificador directo que pueda ocasionar problemas de identificación de individuos, no hemos tenido limitaciones en lo que a protección de datos se refiere. Precisamente, en lo que respecta a la anonimización, los cuasi-identificadores se han generalizado (por ejemplo, la edad se representa en franjas y no con el valor exacto).

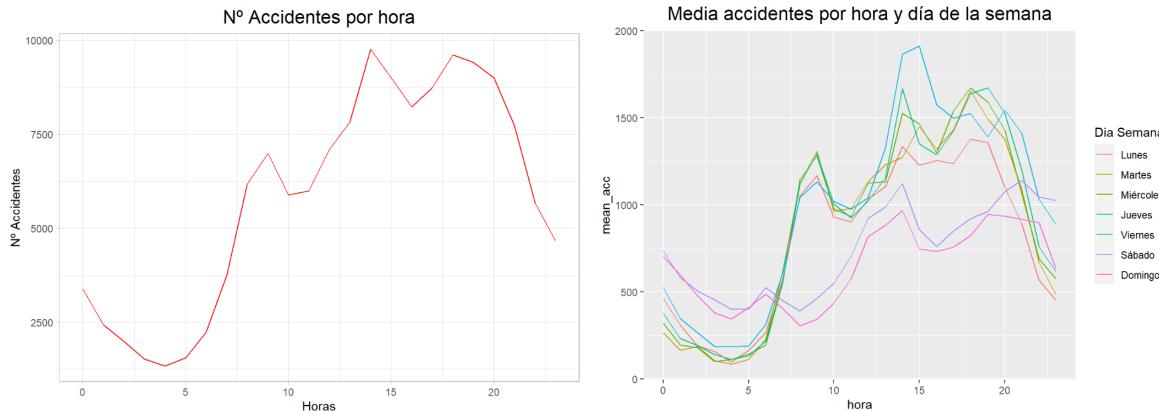
La información relativa a la obtención de los datos se encuentra detallada en los **anexos 1 y 2** del proyecto, en los cuales se explica la **captura** y la **integración** de los datos, respectivamente.

# Desarrollo

*NOTA: Todas las gráficas se adjuntan a mayor tamaño y por orden en el **anexo 3**.*

## Objetivo 1: Estudio por criterios temporales

Uno de los aspectos más relevantes de los accidentes es el momento del día en el que se producen. Por tanto, podemos analizar su frecuencia según la hora del día:

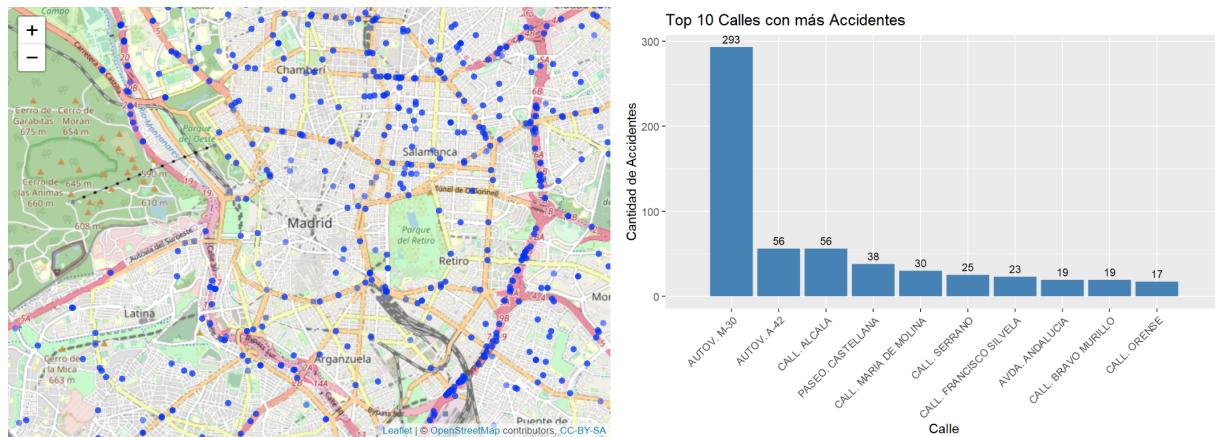


*FIGURA 1 / Distribución de los accidentes por hora. Global (izquierda) y filtrado por día de la semana (derecha)*

Podemos ver picos de siniestralidad a las 09, las 14 y las 17 horas, coincidiendo con las horas punta de entrada y salida de los centros de trabajo o estudio. Este pico es más pronunciado los viernes, cuando la gente que vive fuera de Madrid regresa a sus casas para el fin de semana. Según el periódico El Correo: «El tráfico de un viernes por la tarde llama la atención por el número de coches, pero también por el nivel de ansiedad y de tensión de los conductores. Empieza la fiesta, terminas de trabajar, has hecho planes y vas a todo correr... Cualquiera que salga a la carretera un viernes cuando la gente sale del trabajo lo puede ver».

También hay una lógica disminución de los accidentes por la noche (desde las 21 hasta las 04 horas), que es más pronunciada los días laborables y más ligera los viernes y los sábados. Asimismo, hay un periodo valle entre los dos picos de la tarde, de 14 a 17 horas, que destaca bastante más los fines de semana.

Centrándonos ahora en el pico de los viernes, hemos creado una función con R, utilizando la librería leaflet, para elaborar mapas a partir de nuestras coordenadas. Esta función se utilizará también en próximos objetivos, modificada para reflejar distintas variables (como el tipo de vehículo o el tipo de accidente). En este caso, en la siguiente página mostramos los accidentes que se producen en Madrid a esa hora, junto con un gráfico de frecuencias que refleja las vías con más accidentes:

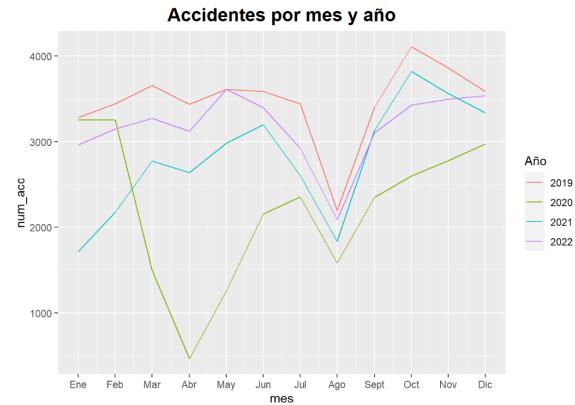


**FIGURA 2** / Mapa de accidentes (izq.) y gráfico de frecuencia de accidentes por vía (der.). Filtrado: viernes a las 15 horas

Podemos ver que gran parte de los accidentes ocurren en la circunvalación M-30, donde se concentran más puntos azules en nuestro mapa.

Más allá del análisis por día de la semana, también hemos considerado relevante analizar la situación por meses. Vemos cómo la accidentalidad desciende en verano y aumenta en los meses de noviembre y diciembre. También aparece en la comparativa el notable descenso de los accidentes en el año 2020: de esto hablaremos en el **objetivo 5**.

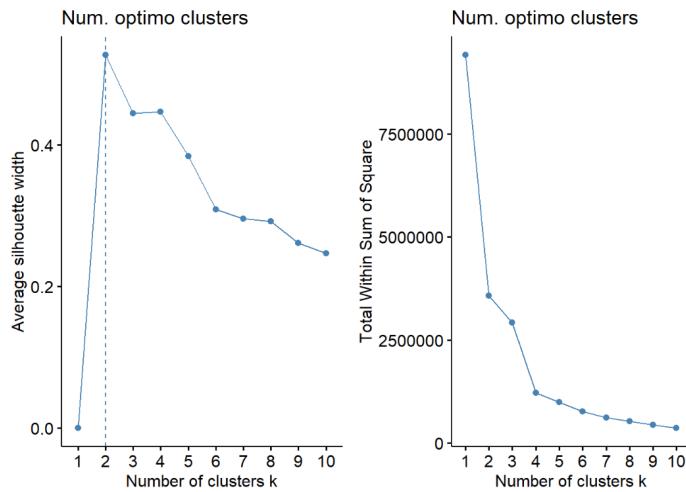
**FIGURA 3** / Distribución de los accidentes por mes del año →



Teniendo datos tan amplios, estimamos necesario aplicar clustering a nuestro *dataset*, para analizar más detalladamente los accidentes según las franjas horarias en las que estos se producen. Tal y como podemos ver en la figura de la siguiente página, a partir de los gráficos de Silhouette y Suma de Cuadrados, tomamos k=4 como número óptimo de clusters. Con este clustering, agruparemos los distritos de la ciudad que presenten un comportamiento similar en lo que a siniestralidad vial se refiere. Esta es la distribución de los distritos en los cuatro clusters:

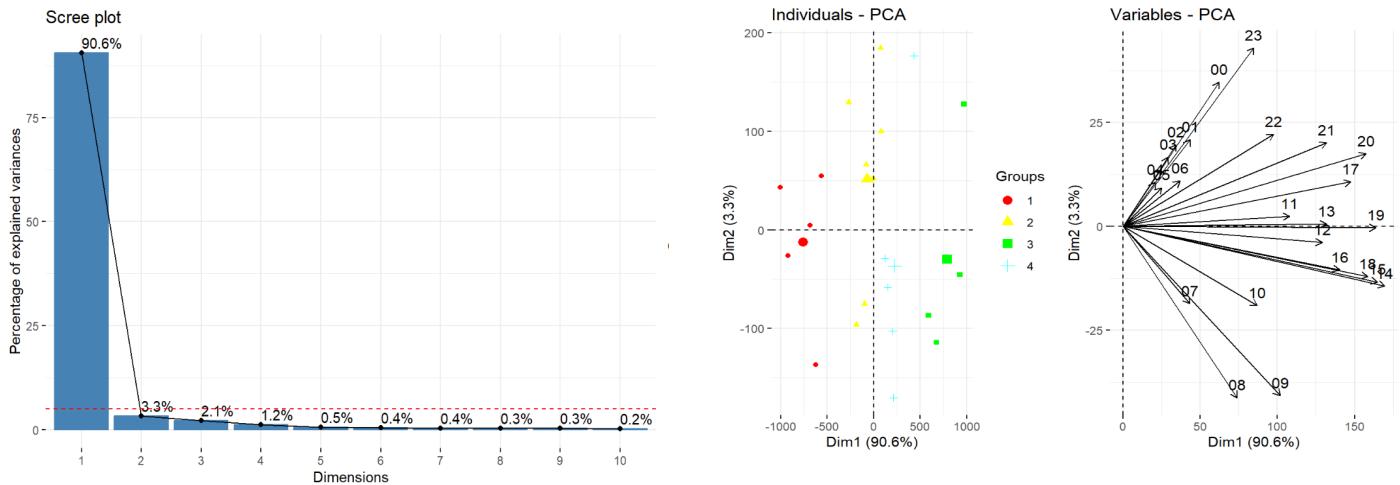
```
## Cluster 1 : BARAJAS MORATALAZ VICÁLVARO VILLA DE VALLECAS VILLAVERDE
## Cluster 2 : ARGANZUELA CENTRO CHAMBERÍ HORTALEZA LATINA TETUÁN USERA
## Cluster 3 : CHAMARTÍN CIUDAD LINEAL PUENTE DE VALLECAS SALAMANCA
## Cluster 4 : CARABANCHEL FUENCARRAL-EL PARDO MONCLOA-ARAVACA RETIRO SAN BLAS-CANILLEJAS
```

**FIGURA 4** / Lista de distribución por clusters de los distritos, siendo k=4



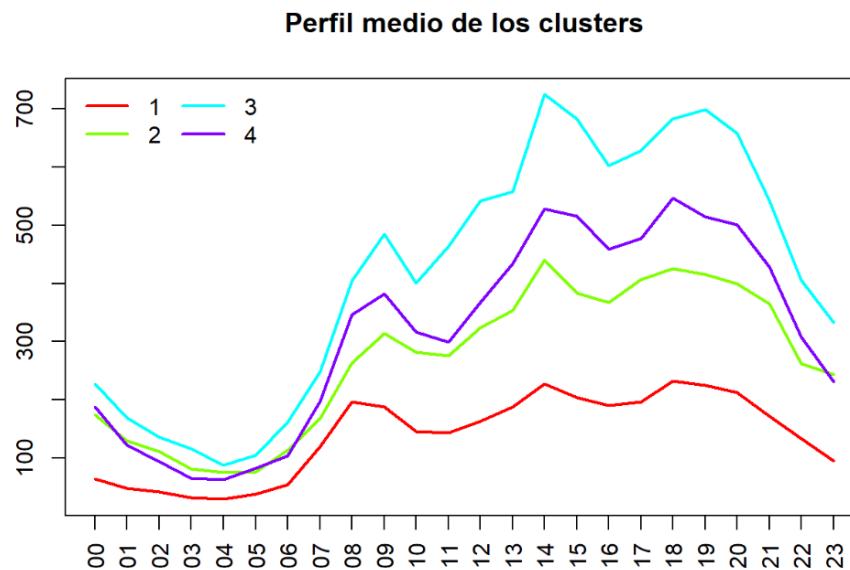
**FIGURA 5** / Gráficos de número de clusters, según Silhouette (izquierda) y Suma de Cuadrados (derecha)

A continuación, agrupando los clusters mediante *kmeans*, realizamos un análisis PCA para observar el comportamiento de los clusters. Al tener unos datos bastante homogéneos, vemos que la primera componente principal explica el 90 por ciento de la variabilidad de nuestra información. Igualmente, adjuntamos a continuación los gráficos de individuos y de variables del análisis PCA:



**FIGURA 6** / Gráficos PCA: porcentaje de varianza explicada, estudio por distritos y estudio por variables

Por último, hemos obtenido el perfil medio de los clusters (cálculo detallado en el [anexo 4](#)):



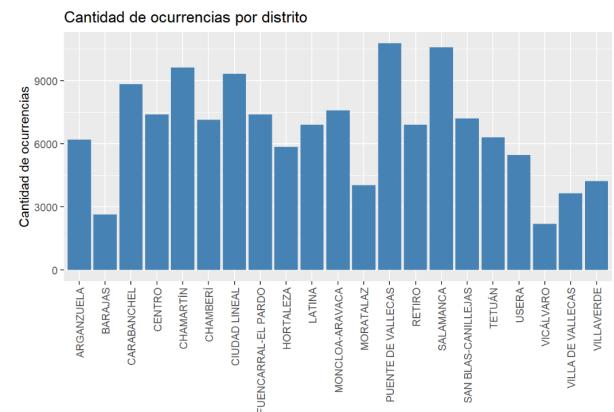
*FIGURA 7 / Perfil medio de los clusters, por hora*

Prácticamente todos los distritos de los clusters presentan un perfil similar en cuanto al número de accidentes entre la medianoche y las 7 de la mañana. Las horas en las que más variabilidad hay es entre las 9 de la mañana y las 9 de la noche. Confirmando lo antes expuesto, se observa mucha variabilidad entre los distritos pertenecientes al cluster 1 y al 3, presentando además este último un pico de accidentes mucho mayor de accidentes a las 8-9 de la mañana, 3 de la tarde y 8 de la noche (horas punta de entrada y salida del trabajo). Tal y como veremos en el próximo apartado en referencia al [objetivo 2](#), este cluster incluye a dos de los tres barrios con más accidentes: Puente de Vallecas y Salamanca. Además, forman parte de este cluster dos barrios que incluyen nodos de transporte con mucho movimiento: en Barajas está el aeropuerto de la ciudad, y en Chamartín se ubica la estación de tren homónima.

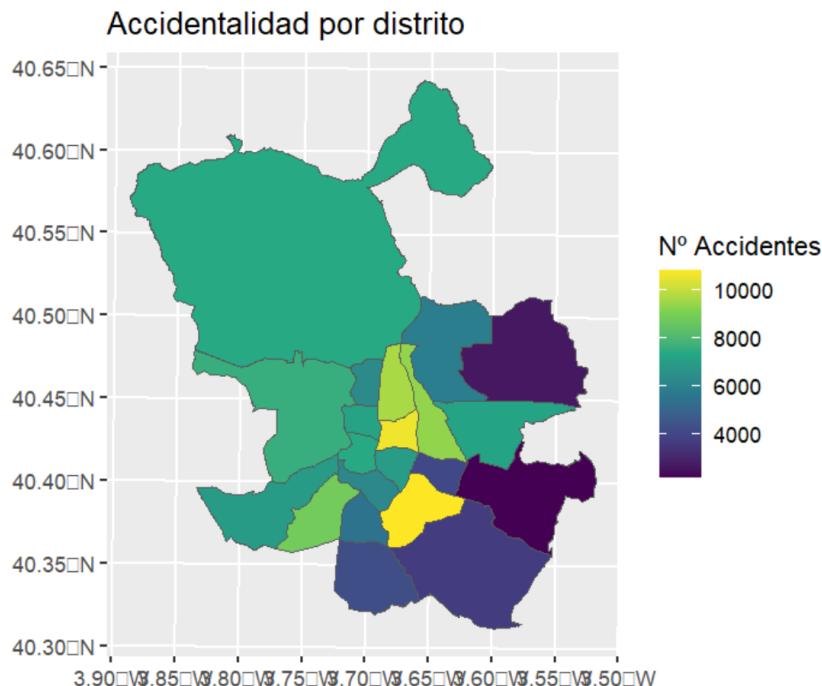
## Objetivo 2: Estudio por ubicación

Estudiaremos en profundidad tres de los distritos con mayor número de accidentes: Salamanca, Puente de Vallecas y Centro. En un primer momento decidimos agrupar los distritos con más accidentes en un cluster, de forma similar al aplicado en el **objetivo 1**, con el objetivo de localizar “zonas de accidentes” más amplias que un solo distrito, ya que gran parte de las vías no pasan solo por un distrito. Buscábamos ver distritos relacionados por estar conectados por las mismas vías. Sin embargo, debido a que solo se incluía una variable, el resultado del *clustering* no nos aporta ninguna información de interés. En cualquier caso, hemos incluido este procesado en el **anexo 4**.

*FIGURA 8 / Frecuencia de accidentes según distrito →*



Esto puede ser más observable si lo visualizamos sobre un mapa de la ciudad. Para ello, adjuntamos un mapa coroplético de la siniestralidad por distrito:



*FIGURA 9 / Mapa coroplético de frecuencia de accidentes por distrito*

En primer lugar, vemos la situación en el céntrico barrio de Salamanca:

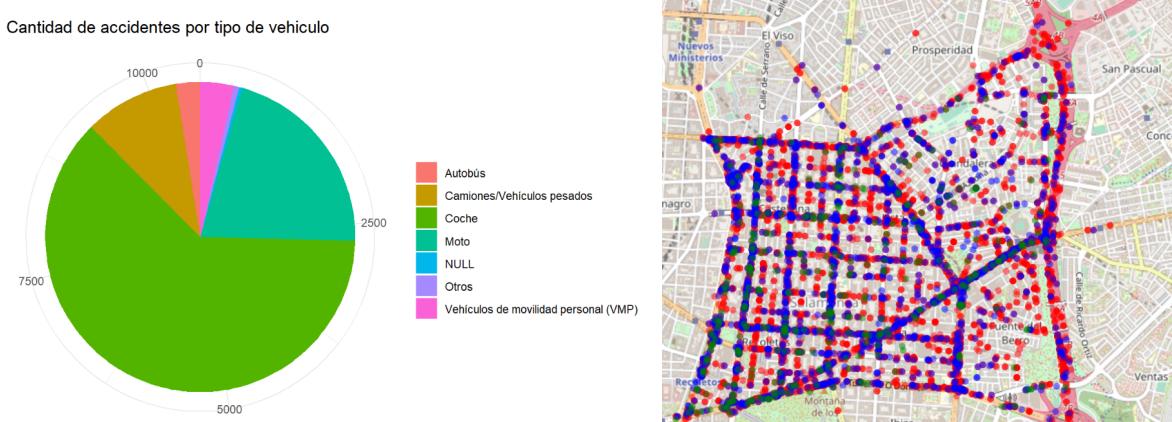


FIGURA 10

**IZQUIERDA:** Gráfico de sectores de frecuencia de accidentes por tipo de vehículo en el distrito de Salamanca  
**DERECHA:** Mapa de accidentes en el distrito de Salamanca: Coches (rojo), Motos (azul) y VMP (verde)

Para los tres análisis, utilizamos la función de mapeado explicada en el objetivo anterior. En este distrito, gran parte de los accidentes son de coche, aunque aproximadamente en uno de cada cinco accidentes hay una moto involucrada. Esto se explica porque Salamanca es un distrito con calles estrechas, muchas de ellas en sentido único, y con un gran número de edificios antiguos, lo cual puede dificultar la circulación de vehículos y aumentar el riesgo de accidentes, especialmente si no se toman las medidas de seguridad adecuadas.

Centrándonos ahora en el distrito de Puente de Vallecas:

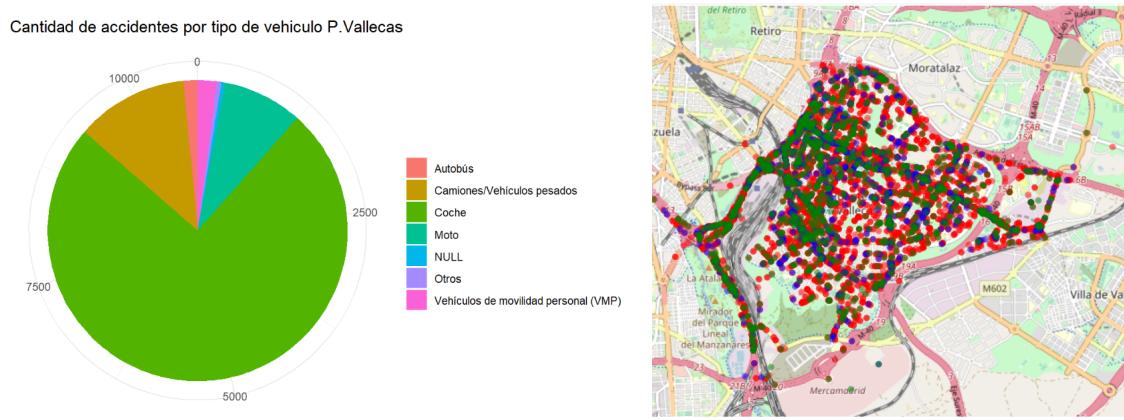
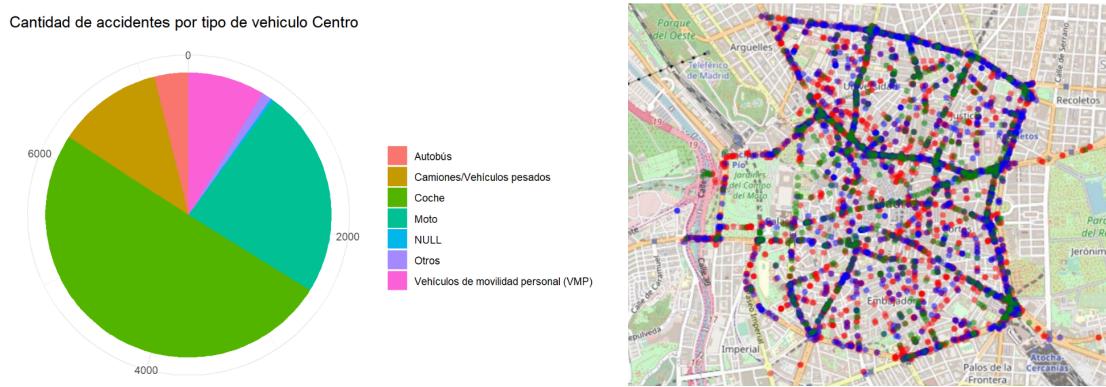


FIGURA 11

**IZQUIERDA:** Gráfico de sectores de frecuencia de accidentes por tipo de vehículo en el distrito de Puente de Vallecas  
**DERECHA:** Mapa de accidentes en el distrito de Puente de Vallecas: Coches (rojo), Motos (azul) y Camiones (verde)

Aquí, la proporción de siniestros de coche es más elevada, ya que en las coordenadas del distrito se incluye la M-30, una de las autovías de circulación de Madrid. También consideramos relevante destacar en el mapa los accidentes de camión, que se concentran en las vías principales y accesos a Mercamadrid (parte inferior del mapa), donde hay mucha concentración de transportistas.

Por último, analizamos el distrito Centro:



**FIGURA 12**

IZQUIERDA: Gráfico de sectores de frecuencia de accidentes por tipo de vehículo en el distrito Centro

DERECHA: Mapa de accidentes en el distrito Centro: Coches (rojo), Motos (azul) y VMP (verde)

En el centro, se mantienen en alza los accidentes de coche, aunque los accidentes de moto y de vehículos de movilidad personal representan un 25% y un 10% respectivamente. Los accidentes se concentran en grandes avenidas, ya que la presencia de vías principales o intersecciones complicadas, puede contribuir a un mayor riesgo de accidentes. También creemos necesario resaltar que en este distrito hay una frecuencia de accidentes de VMP notablemente elevada, debido a que en Madrid no hay una red de carriles bici segregada muy densa (cosa que sí que ocurre en ciudades como Valencia o Barcelona).

## Objetivo 3: Influencia de la meteorología

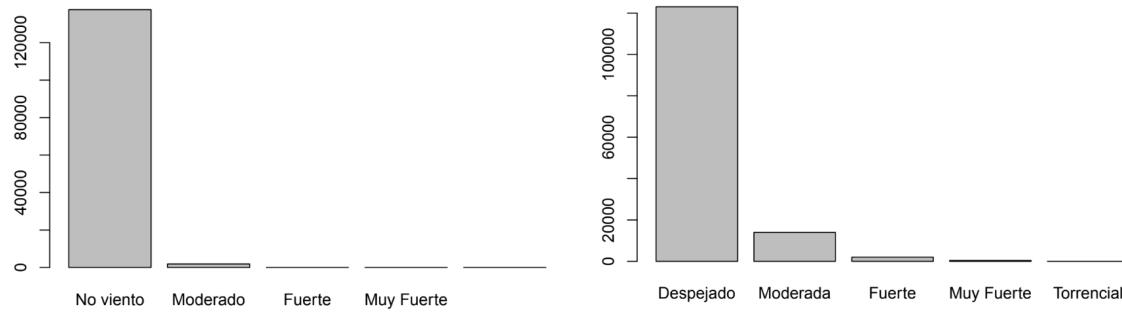
Para poder establecer conexión entre los accidentes y la meteorología, debemos incluir la información relativa a la localización de las cinco estaciones meteorológicas seleccionadas (ver [anexo 2](#)): Aeropuerto, Ciudad Universitaria, Retiro, Cuatro Vientos y Getafe.

En esta base de datos aparecen numerosos datos faltantes, sobre todo en las variables relacionadas con el viento, por lo que ha sido necesario una limpieza previa. Está detallada en el fichero .rmd correspondiente, en el [anexo 4](#).

La clave en el desarrollo de este objetivo ha sido elaborar una función que, mediante la media ponderada de las distancias a las estaciones meteorológicas más cercanas, nos ha permitido obtener las precipitaciones y las rachas de viento de cada uno de los accidentes. Añadiendo estos valores al *dataset* podremos crear análisis más avanzados. Por otra parte, también hemos estimado necesaria una generalización del atributo referente al tipo de vehículo. Hemos agrupado los 39 tipos de vehículo en seis clases más generales: Coche, Moto, Autobús, Vehículo pesado, Vehículo de Movilidad Personal (VMP) y Otros. Se ha utilizado una clasificación similar en el [objetivo 2](#).

Asimismo, hemos discretizado las variables de precipitación y viento en cinco rangos de valores:

- Para el viento: No viento, Moderado, Fuerte, Muy fuerte y Huracanado.
- Para las precipitaciones: Despejado, Moderada, Fuerte, Muy Fuerte y Torrencial.

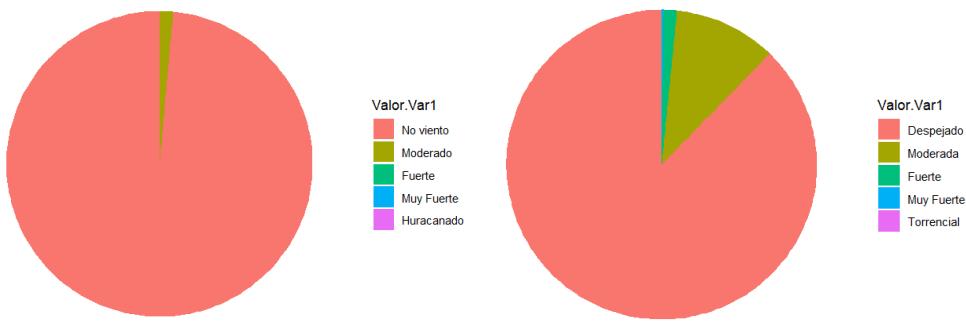


*FIGURA 13 / Frecuencia de accidentes según situación del viento (izquierda) y de las precipitaciones (derecha)*

Elaborando un gráfico de barras para los accidentes en función de dichas variables, observamos que la gran parte de los siniestros ocurren en buenas condiciones meteorológicas.

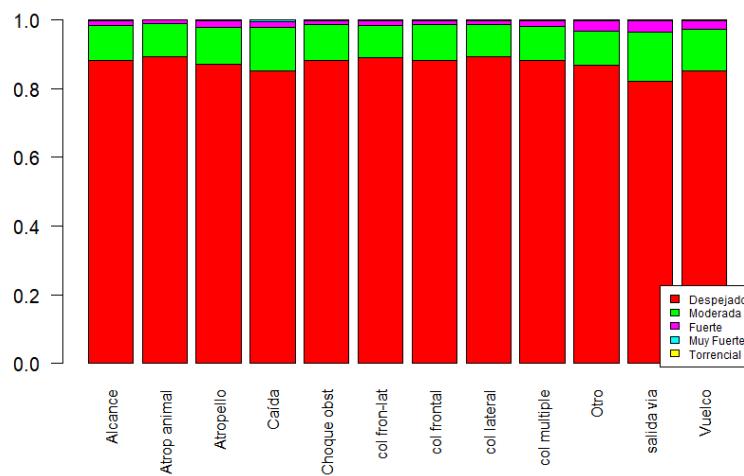
Por lo general, no existen rachas de viento muy fuertes (más de 70 km/h) en la ciudad de Madrid, las cuales son las que podrían afectar a la conducción y/o a la estabilidad del propio vehículo. Por tanto, pensamos que esta es la razón por la que el viento no es un factor significativo en la ocurrencia de accidentes ni en ninguna variable relativa a su clasificación.

Los gráficos de siniestralidad según situación del viento por tipo de vehículo, accidente o vía no nos permiten extraer conclusiones claras, pero en cualquier caso se incluyen en el [anexo 4](#). Tal y como podemos ver en los siguientes gráficos, el número de días de los cuatro años analizados en los que la situación meteorológica ha sido favorable (color rojo) es muy elevado, lo cual refuerza que hayamos descartado los gráficos nombrados anteriormente.



*FIGURA 14 / Gráfico de sectores del porcentaje de días según situación del viento (izq.) y de las precipitaciones (der.)*

Sin embargo, con respecto a la lluvia sí que es algo más visible su influencia en la accidentalidad según su tipo. Tal y como podemos observar, se produce un mayor número de caídas y de salidas de la vía en días con lluvia moderada (color verde). Este tipo de accidentes suele producirse en vehículos en los que el conductor está descubierto, como motos o VMP.



*FIGURA 15 / Frecuencia relativa de accidentes según el tipo de accidente, filtrado por nivel de precipitación*

En conclusión, al contrario de lo que nos imaginamos, no hay una relación entre la cantidad de lluvia y el número de accidentes, ni tampoco con el tipo de vehículo de la persona accidentada. Por tanto, podemos inferir que en Madrid la meteorología no es un factor relevante en el número de accidentes que se producen.

## Objetivo 4: Contraste turismo-moto

En este objetivo nos centraremos en primer lugar en la lesividad de los accidentes, un factor diferencial en los accidentes. Esta variable aparece tanto codificada como descrita, aunque hemos tenido que recodificar algunas respuestas por errores en algunos símbolos. Estas son las clases obtenidas tras la recodificación:

- Asistencia en ambulancia
- Asistencia en centro de salud/mutua
- Asistencia *in situ*
- Atención en urgencias, sin ingreso
- Ingreso inferior a 24 horas
- Ingreso superior a 24 horas
- Fallecido en las primeras 24 horas
- Se desconoce
- Sin asistencia sanitaria

Vamos a establecer comparaciones con respecto a la lesividad, al tipo de vía y al tipo de accidente:

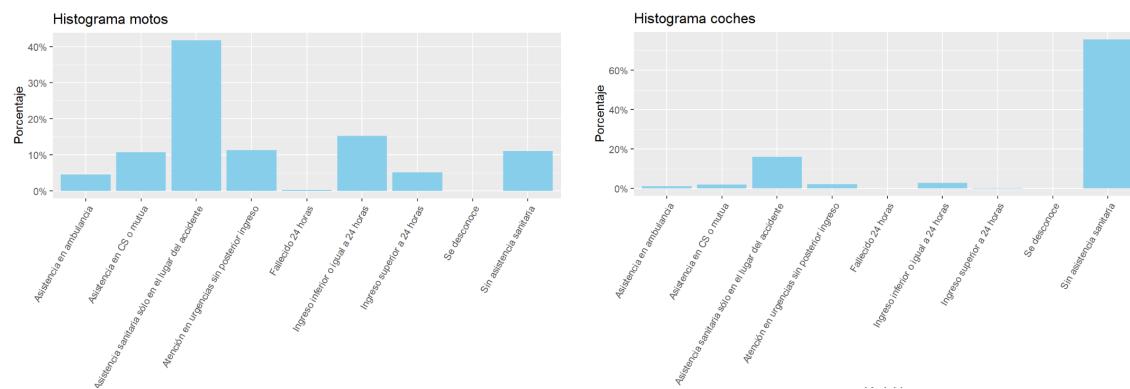


FIGURA 16 / Frecuencia relativa de accidentes según lesividad, para motos (izquierda) y turismos (derecha)

Como podemos observar en la figura anterior, más del 70 por ciento de los accidentes de coche no requieren asistencia sanitaria, mientras que esto solo ocurre en uno de cada diez accidentes de moto. En los segundos, mayoritariamente se resuelven *in situ* (40%), aunque un porcentaje significativo requieren ser ingresados (20% sumando los ingresos breves y los de más de 24 horas).

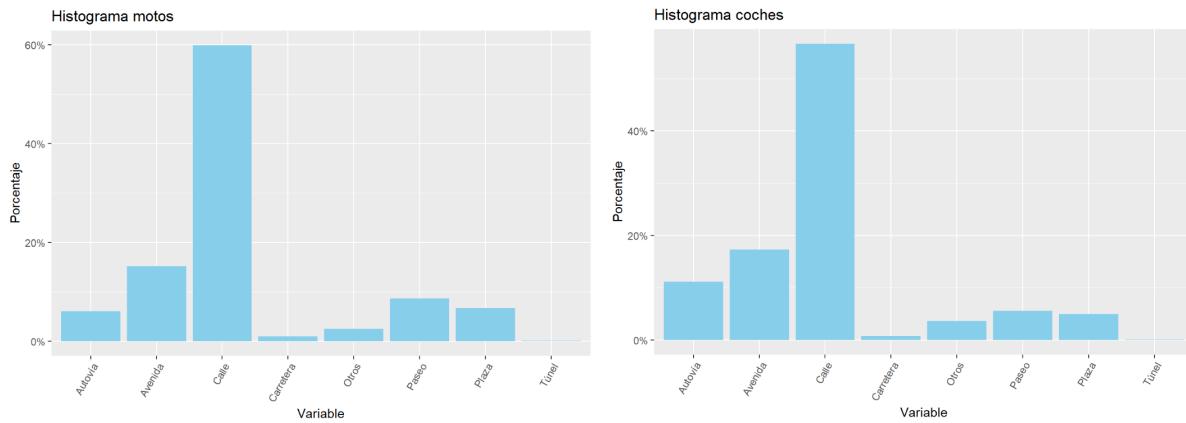


FIGURA 17 / Frecuencia relativa de accidentes según tipo de vía, para motos (izquierda) y turismos (derecha)

Centrándonos ahora en el tipo de vía, aquí no encontramos diferencias relevantes entre ambos vehículos; sin embargo, destacamos que en ambos casos la gran mayoría de accidentes se producen en las calles de la ciudad, que suelen ser más estrechas (y también más numerosas) que otras vías como las avenidas o las autovías, y por tanto es más fácil que se produzca una colisión.

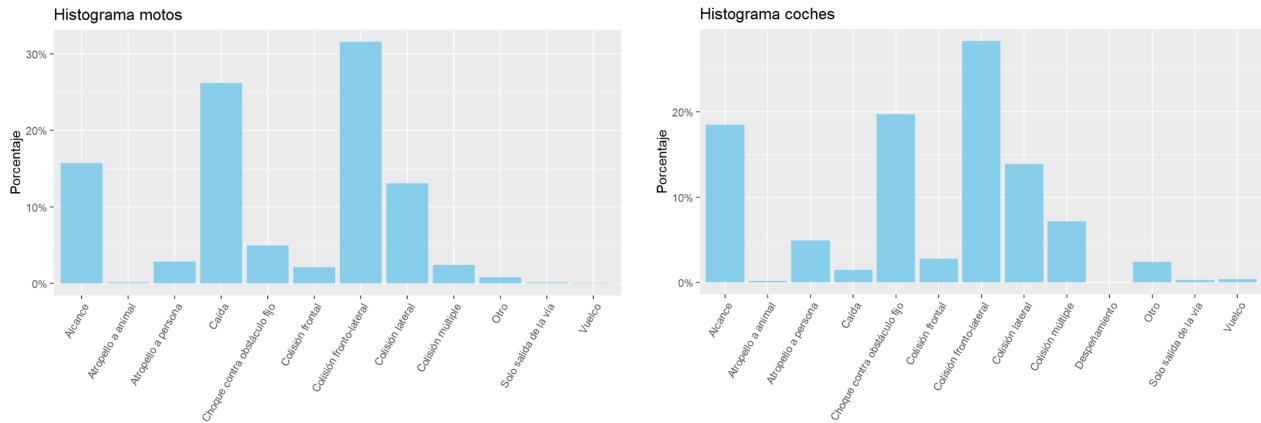
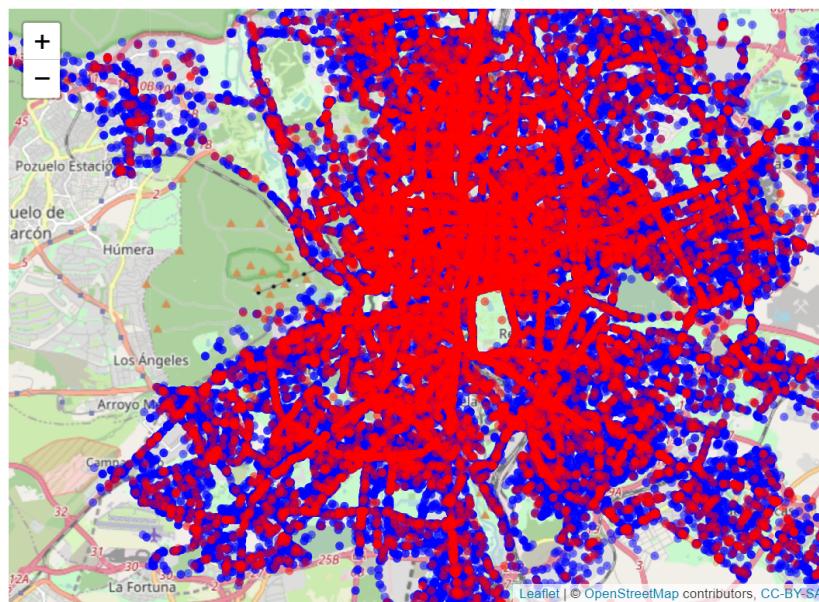


FIGURA 18 / Frecuencia relativa de accidentes según tipo de accidente, para motos (izquierda) y turismos (derecha)

El gráfico relativo a la tipología de los accidentes también presenta una distribución similar, pero con una diferencia muy notable: obviando las colisiones frontal y fronto-lateral, el accidente más común en un vehículo es prácticamente irrelevante en el otro. En las motocicletas, un 25 por ciento de los accidentes son caídas, mientras que en los coches este índice no alcanza el 5 por ciento (es totalmente entendible, ya que es mucho más complicado caerse de un coche que de una moto). Por otra parte, un 20 por ciento de los accidentes de coches son choques contra obstáculos fijos de la vía, mientras que, de nuevo, no se alcanza el 5 por ciento en el otro vehículo (de nuevo, la justificación es lógica; una motocicleta puede esquivar un obstáculo mucho más fácilmente).

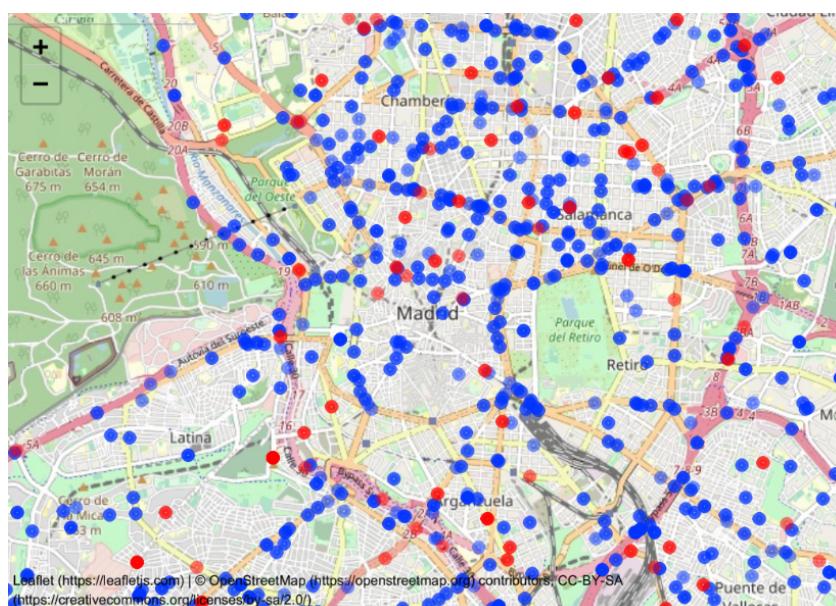
También observamos diferencias notables en cuanto a la ubicación de los accidentes. Como podemos observar en el siguiente mapa (elaborado con la función de mapeado ya vista), los accidentes de moto, marcados en rojo, se concentran en el centro de la ciudad, mientras que los puntos azules, referentes a los accidentes de coche, se dispersan por todo el mapa. Esto se debe a que las motocicletas son vehículos utilizados para desplazamientos más cortos, muchas veces dentro de la propia ciudad, mientras que para desplazarse fuera de Madrid es más habitual el uso del coche.



*FIGURA 19 / Mapa de accidentes en Madrid: Motos (rojo) y Coches (azul)*

## Objetivo 5: Influencia del COVID-19

En primer lugar, utilizando de nuevo la función de mapeado vista en objetivos anteriores, elaboramos un mapa en el que poder observar la diferencia existente en lo que a la siniestralidad vial se refiere. Hemos seleccionado un periodo de tiempo con una movilidad estándar y otro con la movilidad reducida: la segunda quincena de marzo de 2019 y 2020 respectivamente. En estas dos semanas no existe ningún festivo en Madrid más allá del día de San José, el 19 de marzo. Podemos ver en el mapa los accidentes de ambos años por esas fechas:



*FIGURA 20 / Mapa de accidentes en Madrid: 16-30 de marzo de 2019 (azul) y 16-30 de marzo de 2020 (rojo)*

Como podemos ver, reforzando lo observado en el **objetivo 1**, la frecuencia de accidentes es considerablemente menor en 2020. Esto se debe a que en el 15 de marzo de dicho año entraba en vigor el Estado de Alarma en toda España, en el que se decretaba el confinamiento de la población y se limitaba la movilidad.

Ampliando la visión, esta reducción de la movilidad la podemos analizar desde un espectro más amplio, filtrando nuestro *dataset* y tomando todos los meses de marzo y abril de los cuatro años de nuestro estudio. También elaboramos gráficas con un filtro más concreto, centrándonos únicamente en los vehículos más utilizados, y por tanto los que más accidentes sufren: turismos y motocicletas.

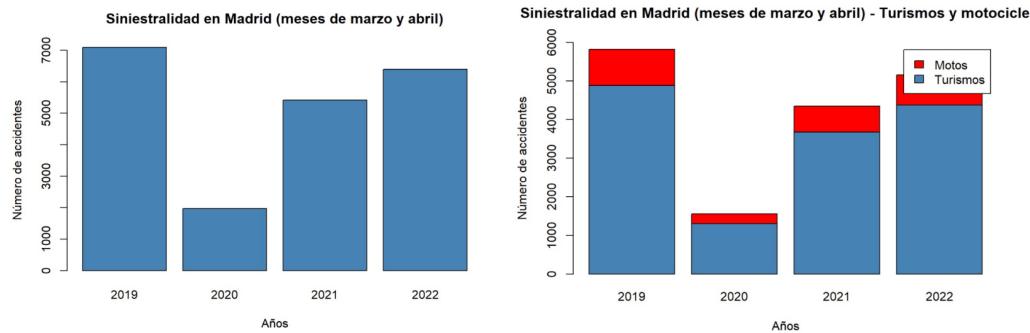


FIGURA 21 / Frecuencia de accidentes en marzo y abril. Global (izquierda) y filtrada por tipo de vehículo (derecha)

Por último, decidimos tomar otro filtro temporal distinto, comparando los meses de julio y agosto de estos años, ya que en España la movilidad es en general significativamente mayor en estos dos meses (en Madrid suele haber más movilidad exterior que interior, ya que los desplazamientos suelen producirse hacia fuera de la ciudad).

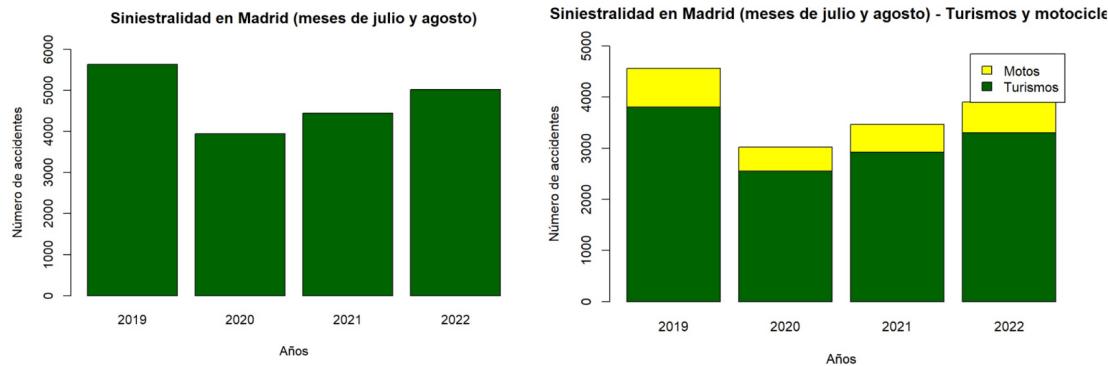


FIGURA 22 / Frecuencia de accidentes en julio y agosto. Global (izquierda) y filtrada por tipo de vehículo (derecha)

Podemos extraer dos conclusiones claras a partir de los gráficos:

- La primera es la reducción de los accidentes en el año 2020: tal y como hemos visto en los mapas, en 2020 se redujo drásticamente la movilidad. Sin embargo, si comparamos los gráficos de primavera y verano, observamos que la siniestralidad de 2020 es algo más elevada en la segunda estación. Esto ocurrió debido a que la situación sanitaria de aquellos meses fue mejor.
- La segunda es la mejora global de la movilidad vial: tanto en primavera como en verano, observamos que no se ha alcanzado el número de accidentes prepandemia (en 2019 hubo más accidentes que en 2022). Podemos entender que todavía no se había recuperado la normalidad total en la sociedad, por lo que, viendo la progresión, es posible que la siniestralidad en 2023 se acerque más, o incluso supere, a la del año 2019.

# Visión de futuro

Este proyecto nos ha servido para desarrollar nuestras competencias en el ámbito de la Ciencia de Datos, trabajando con herramientas tanto teóricas como de software que pueden sernos útiles en el futuro. El principal instrumento utilizado para llevar a cabo nuestros análisis ha sido el lenguaje de programación R, y más concretamente el entorno **R Studio**. Hemos elaborado muchos procesos mediante este entorno, utilizando numerosas librerías:

- Adaptar la base de datos para cada objetivo añadiendo las variables necesarias para cada objetivo: librerías *lubridate* y *dplyr*
- Transformación de las coordenadas de UTM a latitud y longitud: librería *oce*
- Cruce de los datos de tiempo y accidentes creando una función de R que calcula las distancias a las tres estaciones meteorológicas más cercanas: librería *geosphere*
- Ponderación las precipitaciones y viento a partir de la distancia: librería *geosphere*
- Aplicación de PCA: librerías *factominer* y *factoextra*
- Aplicación de clustering y tablas de contingencia: librerías *cluster*, *nbclust* y *cvalid*
- Visualización de los resultados en mapas pictográficos y coropléticos, así como gráficos dinámicos: librerías *ggplot*, *leaflet* y *shiny*
- Creación de una función que localice nuestros accidentes en un mapa de Madrid: librería *geosphere*

De cara a futuros proyectos, pensamos que deberíamos tener una mejor definición de objetivos y una mejor organización de tiempo y tareas, ya que de esta manera hubiéramos alcanzado nuestras metas de una manera más rápida y eficaz. También pensamos que deberíamos de haber incluído más mapas coropléticos que pictográficos.

En definitiva, gracias al proyecto nos hemos dado cuenta de que R es mucho más práctico para la realización de análisis estadísticos y su visualización debido a su amplia variedad de librerías. Junto a Python (lenguaje utilizado para la extracción de datos API REST de la AEMET — ver **anexo 1**), el lenguaje R se convierte en una herramienta clave tanto para futuras asignaturas como para proyectos más allá de la universidad.

## PROYECTO II – Anexo 1

# Captura de datos



**Ciencia de Datos — Curso 2022/2023**

Coral Montes, Adrián Rico, Juan Tomás, Marc Vicedo, Tingting Wu

## Datos utilizados

El conjunto global de nuestros datos se diferencia en tres datasets distintos:

- **Accidentes:** incluye información sobre los accidentes que se han producido en la ciudad de Madrid desde 2018 a 2022, aunque los datos relativos a 2018 los descartamos, debido a que se presentaban con un formato diferente al resto de años. Estos proporcionan atributos sobre cada accidente registrado en la ciudad de Madrid, tales como la localización, momento en que se produjo, tipo de accidente e información relativa al conductor, entre otros. Toda esta información nos podrá ayudar a determinar las principales características definitorias de los percances producidos.
- **Tráfico:** contiene información sobre el flujo de tráfico de la ciudad de Madrid, registrado mediante estaciones y espiras situadas en diferentes puntos del municipio, las cuales nos dicen cuántos vehículos han pasado cada hora por cada estación y en qué sentido lo han hecho; esto nos puede ser útil a la hora de intentar averiguar el motivo de los accidentes, siendo el exceso o falta de flujo un posible motivo.
- **Meteorología:** alberga datos sobre las condiciones meteorológicas registradas en las diferentes estaciones que la Agencia Estatal de Meteorología (AEMET) dispone en la ciudad de Madrid, tales como la temperatura, el viento o precipitaciones, para el mismo período de tiempo en el que tenemos registros de accidentes (2019-2022). A partir de este *dataset* buscamos determinar si el factor meteorológico influye directamente en los accidentes.

También cabe resaltar que en todos los *datasets* existen coordenadas o ubicaciones, lo cual nos facilita considerablemente la unión y el cruce de los datos entre ellas. El amplio espectro de tiempo que utilizamos en nuestro proyecto nos permitirá cruzarlas con una mayor precisión.

Hemos obtenido nuestros datos mediante dos métodos distintos: por una parte, los dos primeros datasets, relativos a los accidentes y el tráfico en la ciudad de Madrid, provienen de **portales de datos abiertos**, que permiten su descarga tanto en formato Excel (XLSX) como en separación por comas (CSV), lo cual los hacía muy accesibles.

Por otro lado, obtuvimos los datos relativos a la meteorología accediendo a la **interfaz API** proporcionada por la AEMET. En este caso, obtuvimos los datos en formato JSON separados por estaciones meteorológicas, y con un programa en Python los convertimos a formato CSV. Una vez completada la conversión del formato de los datos, los unimos en un único *dataset* con R, añadiendo datos a mano para utilizarlos más tarde, como por ejemplo las coordenadas de las estaciones meteorológicas, obtenidas en la página web de la AEMET.

Dentro de los criterios aplicados para la selección de las fuentes necesarias para la realización del proyecto, destaca la **completitud de los datos**. Nos resultaban de interés aquellas fuentes en las que se disponía de información relativa a distintos períodos de tiempo, y que al mismo tiempo se pudieran relacionar con otras características presentes para disponer de un marco abierto de opciones posibles a utilizar. Al finalizar la búsqueda, nos decantamos por los datos de accidentes de tráfico que publicó el Ayuntamiento de Madrid en su página de datos abiertos, ya que estos ofrecían datos de un amplio rango de años. Esto fue lo que nos hizo decidirnos por Madrid frente a otras ciudades españolas.

---

Aquí adjuntamos enlaces a nuestras fuentes:

- Accidentes: <https://datos.gob.es/es/>
- Tráfico: <https://datos.madrid.es/portal/site/egob>
- Meteorología: <https://opendata.aemet.es/>

## Análisis preliminar de fuentes

Con carácter previo a elaborar un análisis exploratorio de nuestros datos, estimamos necesario entender en profundidad los campos que componen nuestras bases de datos.

De la base de datos relativa a los **accidentes** que se han producido en la ciudad de Madrid tenemos los siguientes campos:

A	B	C	D	E	F	G	H	I	J	K
num_expediente	fecha	hora	localización	numero	cod_distrito	distrito	tipo_accidente	estado_meteorologico	tipo_vehiculo	tipo_persona
2018S01784; 04/02/2019	04/02/2019	9:10:00	CALL. ALBERT	1	1	CENTRO	Colisión later	Despejado	Motocicleta	> Conductor
2018S01784; 04/02/2019	04/02/2019	9:10:00	CALL. ALBERT	1	1	CENTRO	Colisión later	Despejado	Turismo	Conductor
L	M	N	O	P	Q	R	S			
rango_edad	sexo	cod_lesividad	lesividad	coordenada_x	coordenada_y	utm	positiva_alco	positiva_d		
De 45 a 49 años Hombre		7	Asistencia sanitaria	440068,05	4475679,17	N		NULL		
De 30 a 34 años Mujer		7	Asistencia sanitaria	440068,05	4475679,17	N		NULL		

Los campos en las columnas de la A a la G, así como los referentes a las coordenadas (P, Q), son variables que se presentan clave para el posterior cruce de datos de distintas fuentes, porque mediante la fecha, hora, localización, coordenadas... podríamos inferir en qué punto de la ciudad y a qué hora se produjeron los accidentes, y por tanto, sabríamos las condiciones meteorológicas que hubo ese mismo día.

Muchas de las variables que presenta este *dataset* son categóricas, referentes por ejemplo al distrito (G) o a su respectivo código numérico asociado (F). Otros presentan una codificación más compleja, como puede ser el tipo de vehículo manejado (J), o más sencilla, como es el caso del tipo de persona involucrada en el siniestro (K), que solo puede tomar los valores “Conductor” o “Peatón”.

Por otra parte, aunque la base de datos relativa a la **meteorología** no ofrece el tiempo por horas, sí que presenta una gran cantidad de variables numéricas acompañadas por las horas en las que las condiciones se maximizan o minimizan. A su vez, la columna I de la base de datos de los accidentes ofrece una breve descripción categórica del tiempo en el momento del accidente, lo cual nos puede ayudar a profundizar en el análisis de la relación entre la situación meteorológica y los accidentes. Adjuntamos en la siguiente página los campos correspondientes a la base de datos de meteorología:

A	B	C	D	E	F	G	H	I
	fecha	nombre	tmed	prec	tmin	horatmin	tmax	horatmax
1560	09/04/2019	MADRID CIUI	9,4	NA	4	5:30	14,7	16:30
127	07/05/2019	MADRID AERC	17,7	NA	10,5	5:44	24,9	16:01
135	15/05/2019	MADRID AERC	21,5	NA	9,1	5:29	33,9	14:57

J	K	L	M	N	O
dir	velmedia	racha	horaracha	latitud	longitud
0	99	4,4	13,3 Varias	438594,3	4478141,5
1	22	6,9	16,4	452902,2	4479702,9
7	35	2,2	9,7	452902,2	4479702,9

Para crear la base de datos de **tráfico** hemos cargado los ficheros CSV mes a mes del portal de datos abiertos de Madrid (datos.madrid.es). Hemos juntado los ficheros de cada mes en uno solo, y eliminado las filas con valores nulos en todas sus columnas, ya que suponemos que se trataba de un error al cargar los ficheros en R Studio.

Finalmente, nuestro fichero de tráfico final ha quedado de la siguiente manera: en la primera columna se indica la fecha, la segunda es la estación que ha registrado el tráfico y la tercera muestra el sentido y la franja horaria. El resto de columnas indican las horas. El sentido se refleja con este código:

- 1- → Sentido 1. Datos tomados de 1:00 a 12:00
- 1= → Sentido 1. Datos tomados de 13:00 a 24:00
- 2- → Sentido 2. Datos tomados de 1:00 a 12:00
- 2= → Sentido 2. Datos tomados de 13:00 a 24:00

Adjuntamos aquí la distribución de los atributos de este *dataset*:

A	B	C	D	E	F	G	H	I	J	
	FEST	FSEN	HOR1	HOR2	HOR3	HOR4	HOR5	HOR6	HOR7	
01/05/2022	ES01	1-		550	459	324	284	242	222	129
01/05/2022	ES01	1=		619	640	446	747	1026	948	1048
01/05/2022	ES01	2-		597	495	337	301	310	335	187
01/05/2022	ES01	2=		812	766	694	848	1017	900	875

Las ubicaciones de las estaciones y el sentido las hemos guardado en otro fichero distinto, para facilitar su identificación a posteriori. En este fichero reflejamos los siguientes atributos: la estación, con su ubicación y un código arbitrario; sus coordenadas, el sentido reflejado en el *dataset* de tráfico, y la orientación a la que este corresponde (por ejemplo, S-N sería de sur a norte, y O-E sería de oeste a este). Finalmente, este fichero no ha sido utilizado en nuestros análisis.

## PROYECTO II – Anexo 2

# Integración de datos



**Ciencia de Datos — Curso 2022/2023**

Coral Montes, Adrián Rico, Juan Tomás, Marc Vicedo, Tingting Wu

## Nuevo enfoque

Somos conscientes de que existen estudios similares, aunque estos no se centran en lugares tan concretos como el un área metropolitana, ya que suelen ser a mayor escala. Normalmente, estos se suelen centrar a nivel estatal o autonómico; asimismo, por otro lado, también hay análisis que se centran únicamente en accidentes con víctimas mortales. Adjuntamos en [este enlace](#) el estudio estadístico que elaboró la Dirección General de Tráfico (DGT) junto con la Fundación Mutua a este respecto.

Nuestro estudio presenta como novedades principales, más allá de una dimensión temporal más amplia (los estudios analizados suelen ser anuales), la introducción de un mayor detalle del factor meteorológico, observando la relación que pueden tener la temperatura, viento, precipitación y demás factores de este tipo con el accidente. También permitimos medir la relación de la cantidad de tráfico existente en el momento del accidente con las características del conductor o conductores afectados.

---

## Análisis exploratorio

Pasamos a exponer los análisis, preprocesados y limpiezas realizados en nuestros *datasets* con el objetivo de integrar y transformar los datos, buscando que nos aporten utilidad.

En primer lugar, juntamos los datos de **meteorología** obtenidos a partir de la API de AEMET de todas las estaciones de Madrid. Al observar nuestras bases de datos, un detalle que captó nuestra atención fue la cantidad de valores faltantes, por lo que decidimos realizar un análisis previo. Entonces localizamos estos valores en sus respectivas variables, viendo el porcentaje que estos representaban. Para suprimirlos, usamos la función de R ‘md.pattern’, que analiza los patrones de los valores faltantes por observación, y eliminamos las observaciones que contenían un porcentaje alto de estos, que resultaron ser las referentes al viento.

A continuación, añadimos las coordenadas de cada estación meteorológica para que el cruce con la base de datos de accidentes fuera más preciso. Intentamos separar los datos por estación para poder observar si la variación entre el dato recogido en cada estación para cada día era lo suficientemente distinto como para no considerar diferentes estaciones meteorológicas en Madrid y poder usar solo una; sin embargo, observamos que esto no era

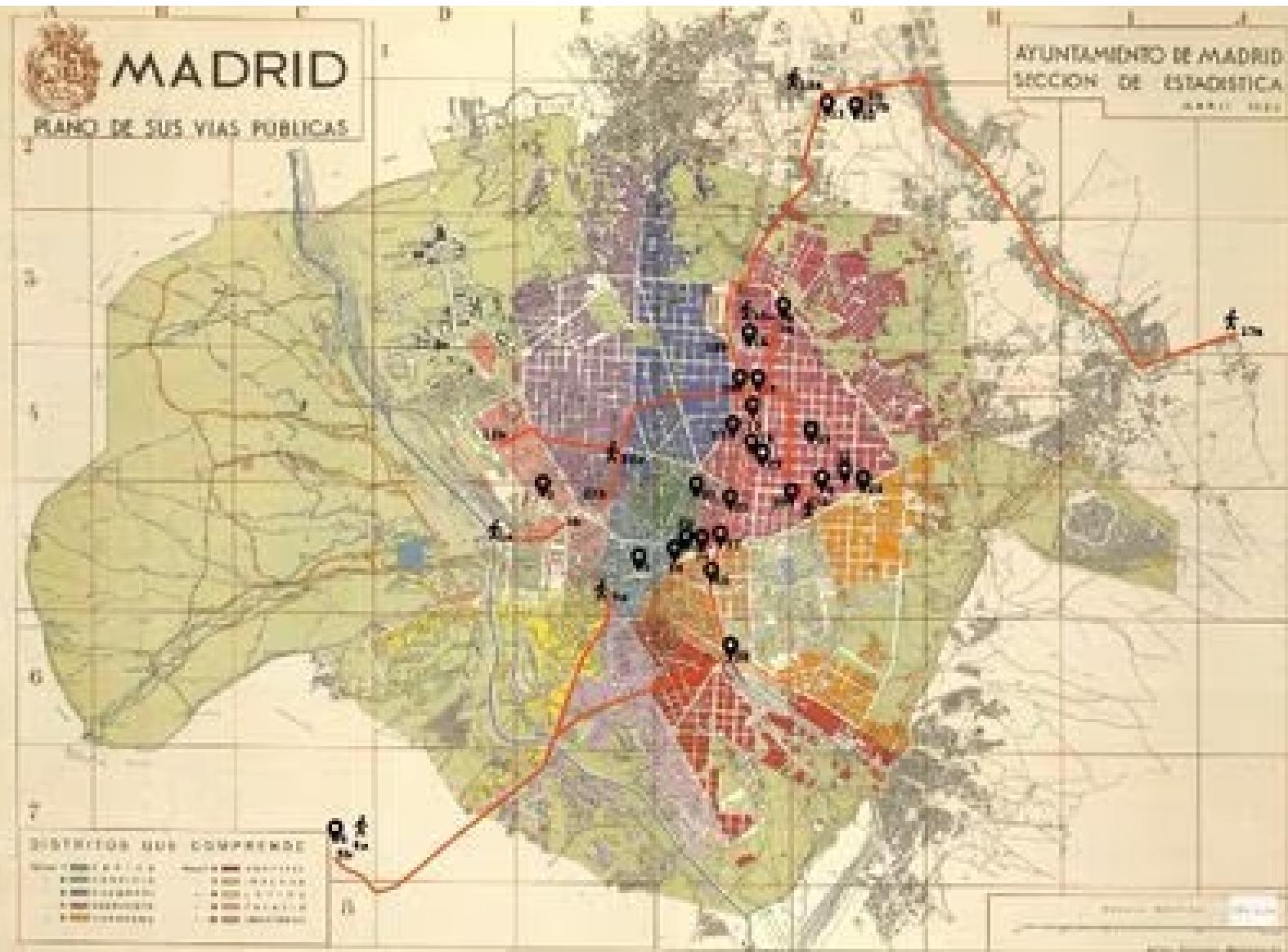
posible, por lo tanto seleccionamos las estaciones más próximas a los accidentes para analizar el accidente con las condiciones meteorológicas más exactas posibles. Las estaciones seleccionadas fueron las de Retiro, Ciudad Universitaria, Cuatro Vientos, Aeropuerto y Getafe.

Centrándonos ahora en la base de datos relativa al **tráfico**, tuvimos que juntar muchos conjuntos de datos, ya que esta información venía separada mes por mes en sus respectivos ficheros CSV. Una vez tuvimos la base de datos completa, tras una primera revisión observamos que había muchos valores numéricos imposibles como 9999 o -99999. Al considerar razonablemente que era muy improbable que pasaran 9999 coches en una hora por un mismo sitio, decidimos convertirlos a valores faltantes; también eliminamos de la base de datos las filas que no contenían ningún dato. Tras este primer preprocesado de los datos, volvimos a analizarlos y observamos que seguía habiendo valores muy extremos; por tanto, normalizamos las variables de las horas utilizando el Z-score, intentando localizar anomalías. Ahora, siguiendo la distribución de Gauss, la práctica totalidad de los valores se encuentra dentro del intervalo [media +/- 3\*desv\_típica], y por tanto recorrimos los valores de todas las columnas y cambiamos las observaciones fuera de este intervalo por NA. En último lugar, también cambiamos los valores de 0 por NA, ya que esto indica que la estación no estaba en funcionamiento a esa hora, y no que no existiese tráfico en dicho momento.

En cuanto a la base de datos relativa a los **accidentes** en el área metropolitana de Madrid, unificamos los cuatro ficheros CSV que teníamos, referentes a los años 2019, 2020, 2021 y 2022 en un único fichero. Vimos que había muchos registros de un mismo accidente, ya que aparecía una fila por cada persona involucrada (tanto conductores como peatones); viendo esto, decidimos filtrar los datos dejando solo el registro relativo a los conductores. Para no perder información sobre el accidente, añadimos una columna referente al número de personas involucradas en dicha situación. También hemos creado “subconjuntos” de datos a partir del conjunto global, en los que solamente se incluyen los accidentes que afectan a un tipo de vehículo determinado (motocicletas, bicicletas, coches...), para así poder contrastarlos y cumplir los objetivos correspondientes. Por último, añadimos dos columnas nuevas a la base de datos, una con el nombre del día de la semana y otra que indica en respuesta binaria si dicho día era festivo. También hemos añadido columnas relativas a los metadatos extraíbles de la fecha (día de la semana, nombre del mes, etc.) mediante la librería de R ‘lubridate’, buscando reflejar las diferencias en los patrones de siniestralidad comparando diferentes intervalos temporales.

# **PROYECTO II – Anexo 3**

# Figuras gráficas



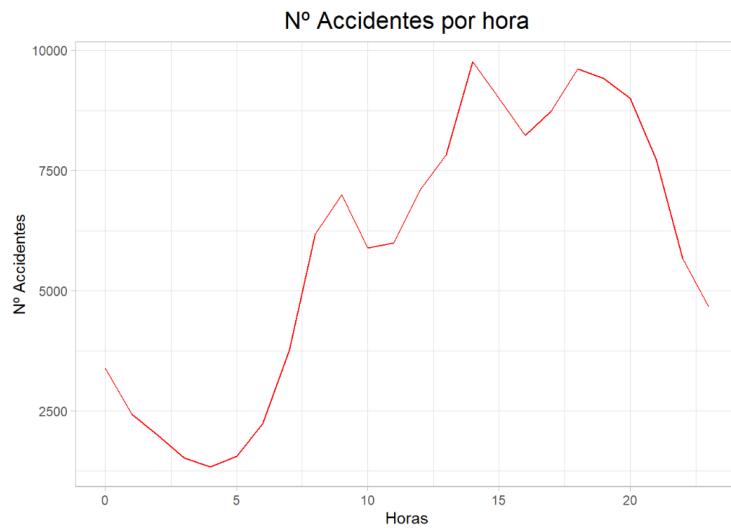
**Ciencia de Datos – Curso 2022/2023**

Coral Montes, Adrián Rico, Juan Tomás, Marc Vicedo, Tingting Wu

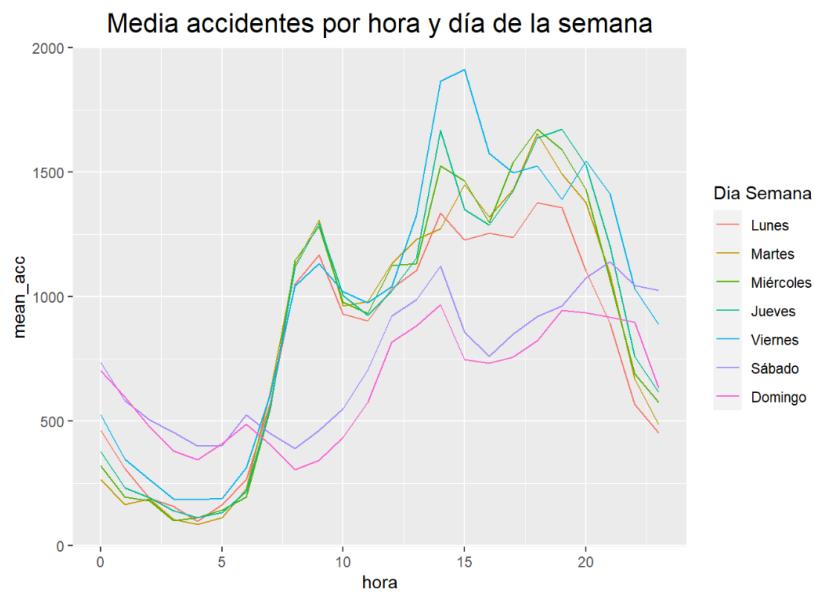
En este anexo adjuntamos todos los gráficos, diagramas y mapas descritos en los objetivos, para su mejor visualización. En el cuerpo de la memoria se ha reducido el tamaño de las figuras para respetar el límite de extensión de páginas que se ha establecido.

## Objetivo 1: Estudio por criterios temporales

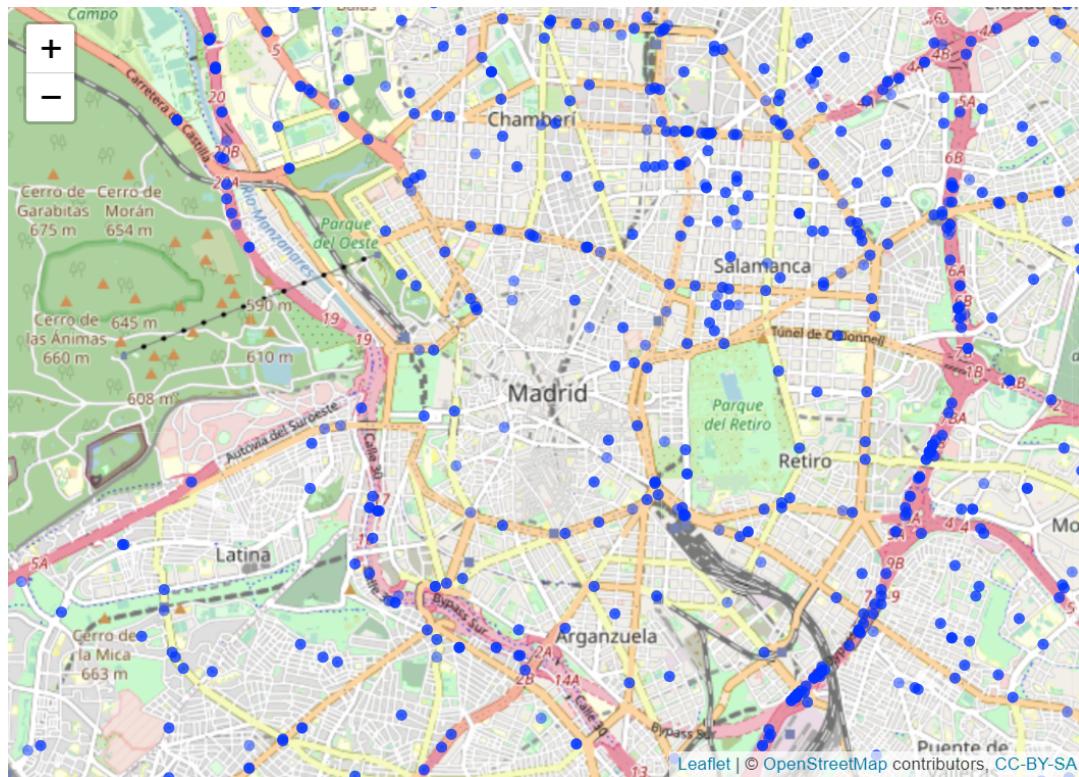
**FIGURA 1A:** Distribución global de los accidentes por hora



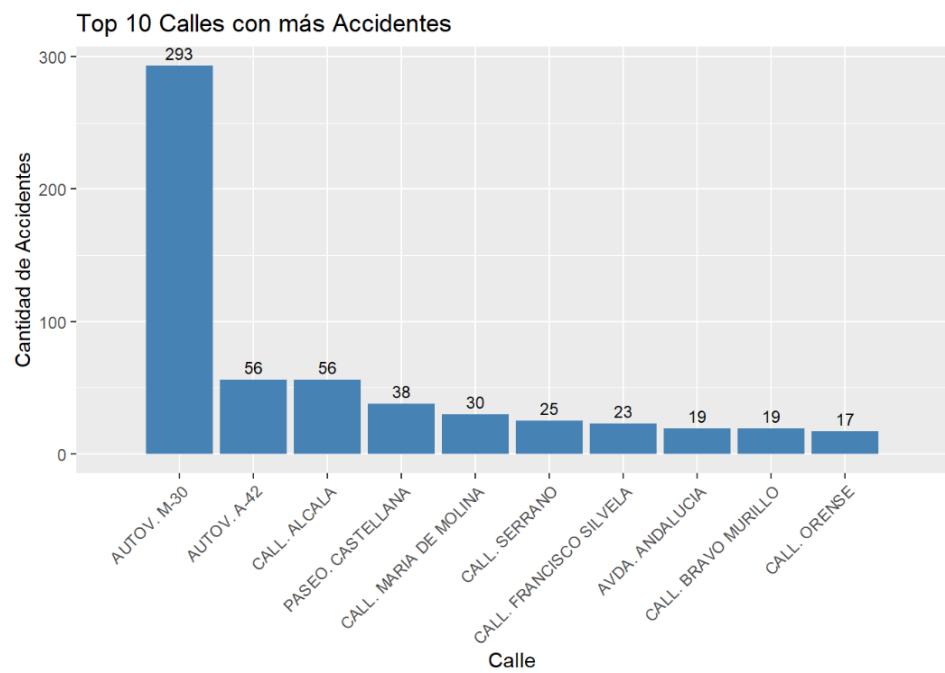
**FIGURA 1B:** Distribución de los accidentes por hora, filtrada por día de la semana



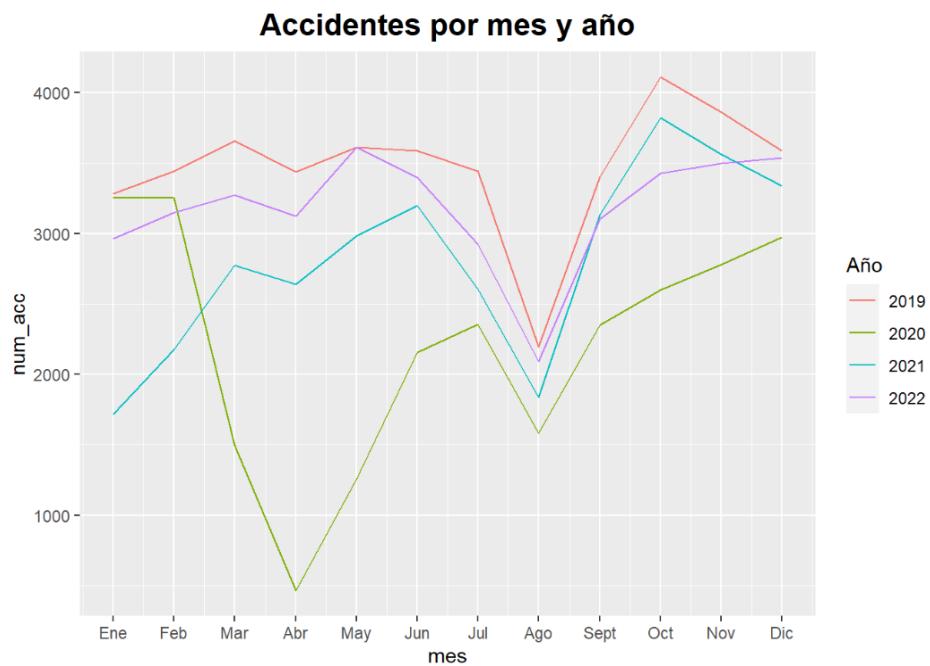
**FIGURA 2A:** Mapa de accidentes, filtrado: viernes a las 15 horas



**FIGURA 2B:** Gráfico de frecuencia de accidentes por vía, filtrado: viernes a las 15 horas



**FIGURA 3:** Distribución de los accidentes por mes del año

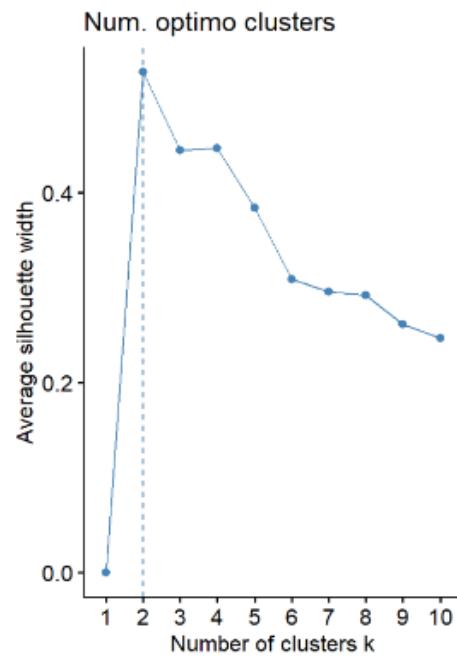


**FIGURA 4:** Lista de distribución por clusters de los distritos, siendo k=4

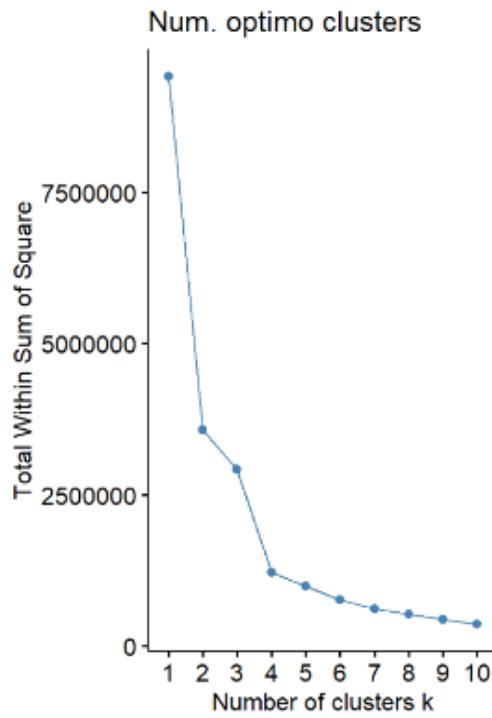
---

```
## Cluster 1 : BARAJAS MORATALAZ VICÁLVARO VILLA DE VALLECAS VILLAVERDE
## Cluster 2 : ARGANZUELA CENTRO CHAMBERÍ HORTALEZA LATINA TETUÁN USERA
## Cluster 3 : CHAMARTÍN CIUDAD LINEAL PUENTE DE VALLECAS SALAMANCA
## Cluster 4 : CARABANCHEL FUENCARRAL-EL PARDO MONCLOA-ARAVACA RETIRO SAN BLAS-CANILLEJAS
```

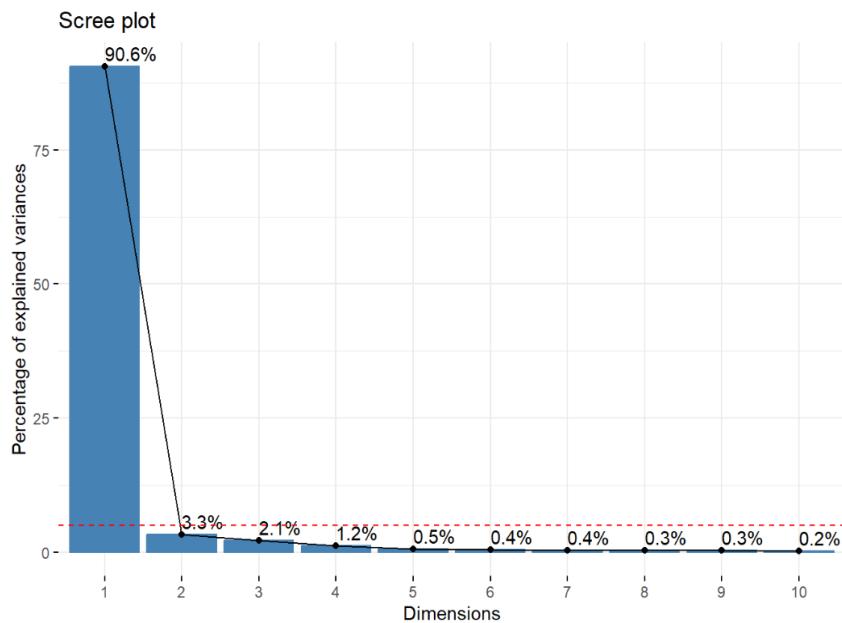
**FIGURA 5A:** Gráfico de número de clusters, según Silhouette



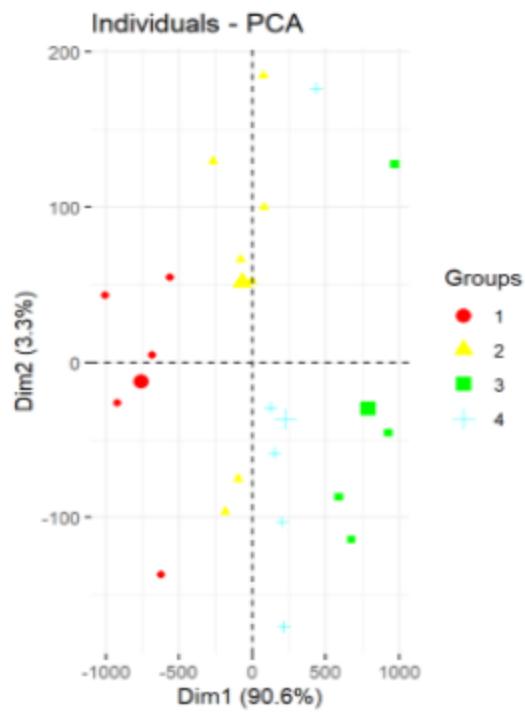
**FIGURA 5B:** Gráfico de número de clusters, según Suma de Cuadrados



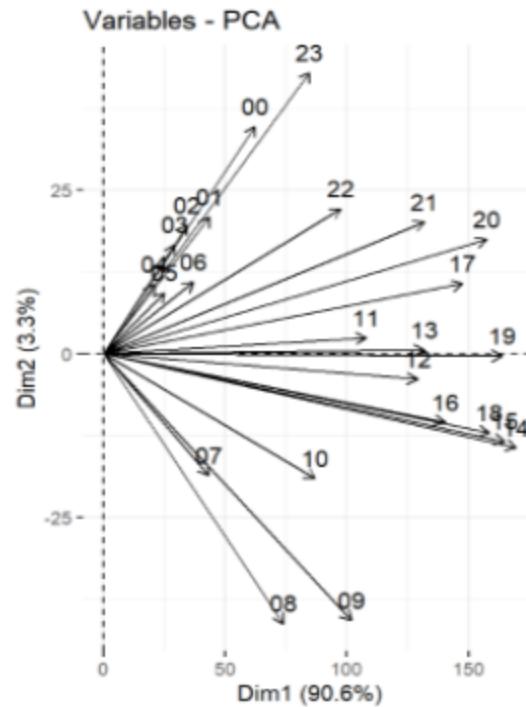
**FIGURA 6A:** Gráfico PCA — Porcentaje de varianza explicada



**FIGURA 6B:** Gráfico PCA — Estudio por distritos

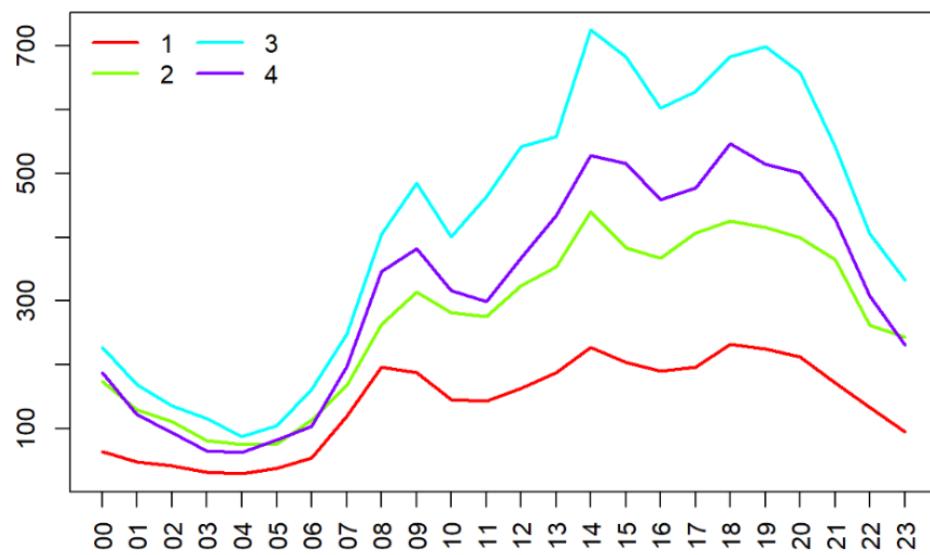


**FIGURA 6C:** Gráfico PCA — Estudio por variables



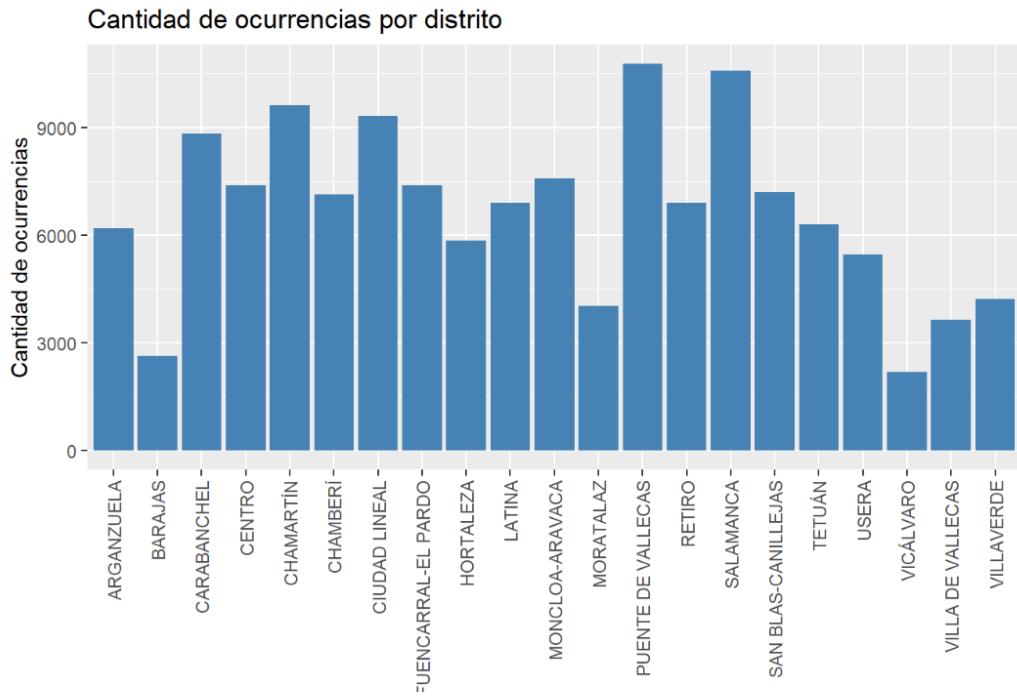
**FIGURA 7:** Perfil medio de los clusters, por hora

Perfil medio de los clusters

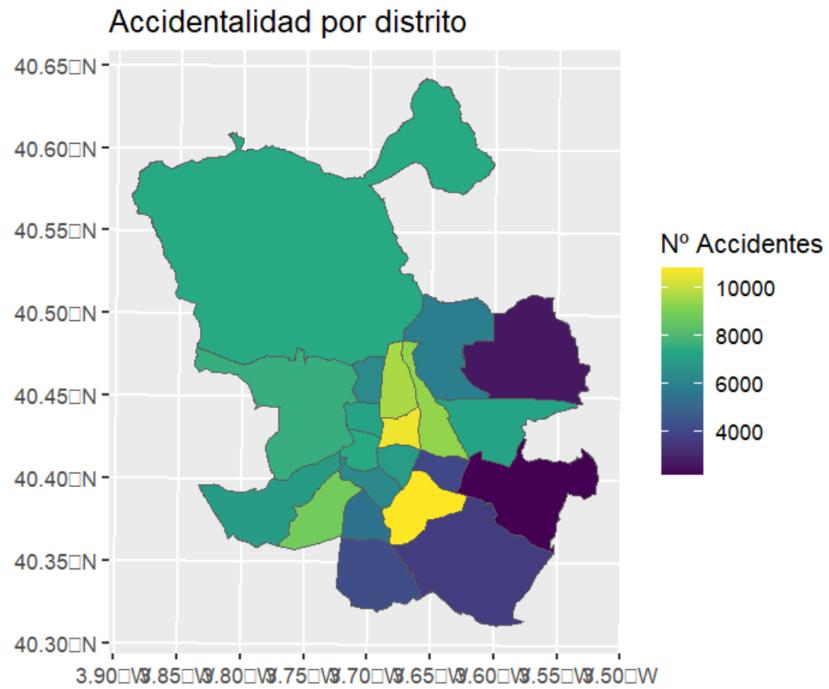


## Objetivo 2: Estudio por ubicación

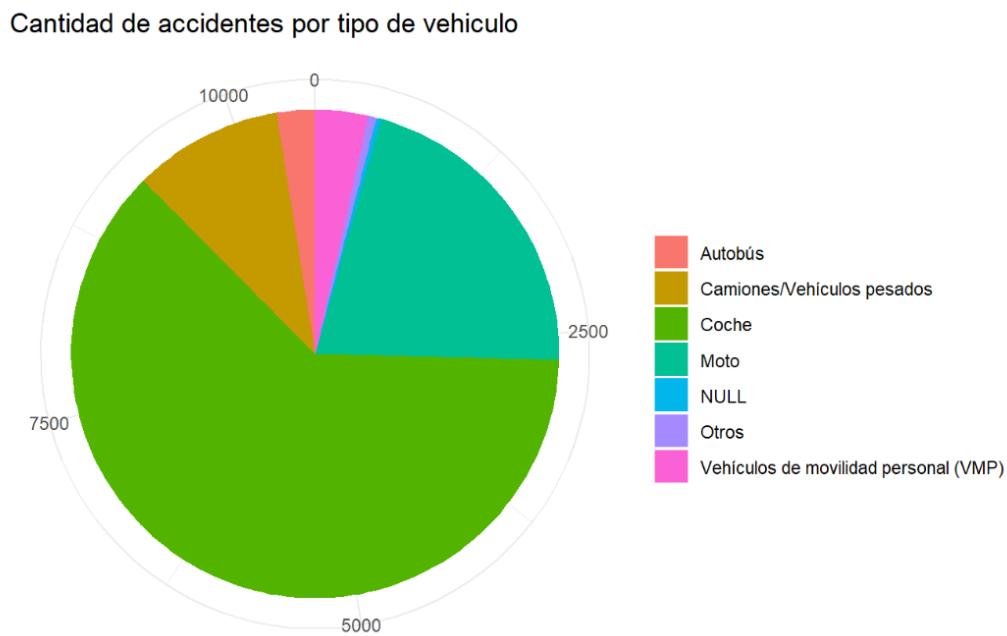
**FIGURA 8:** Gráfico de frecuencia de accidentes según distrito



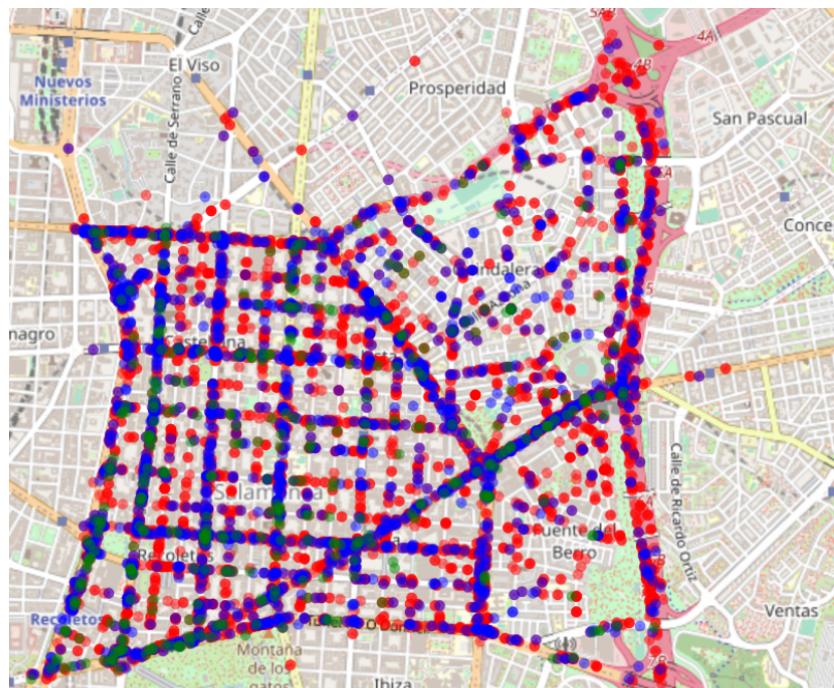
**FIGURA 9:** Mapa coroplético de frecuencia de accidentes según distrito



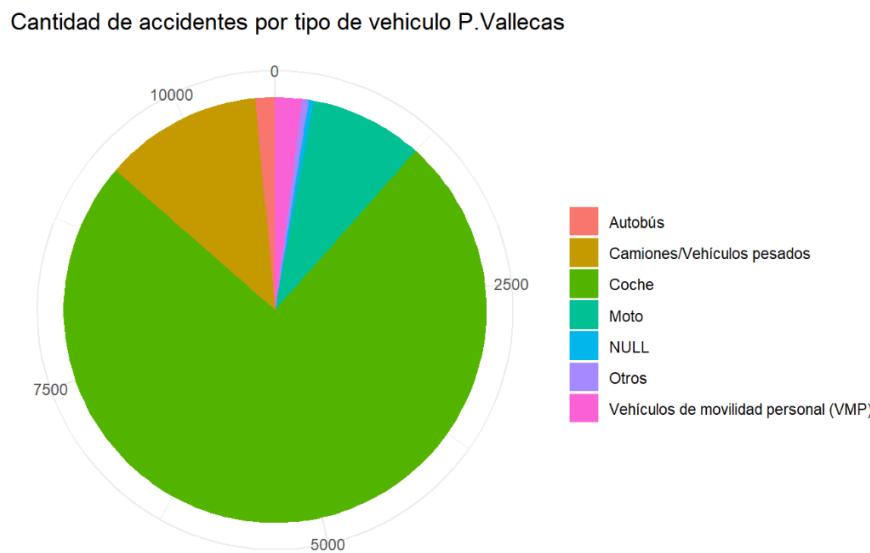
**FIGURA 10A:** Gráfico de sectores de frecuencia de accidentes por tipo de vehículo en el distrito de Salamanca



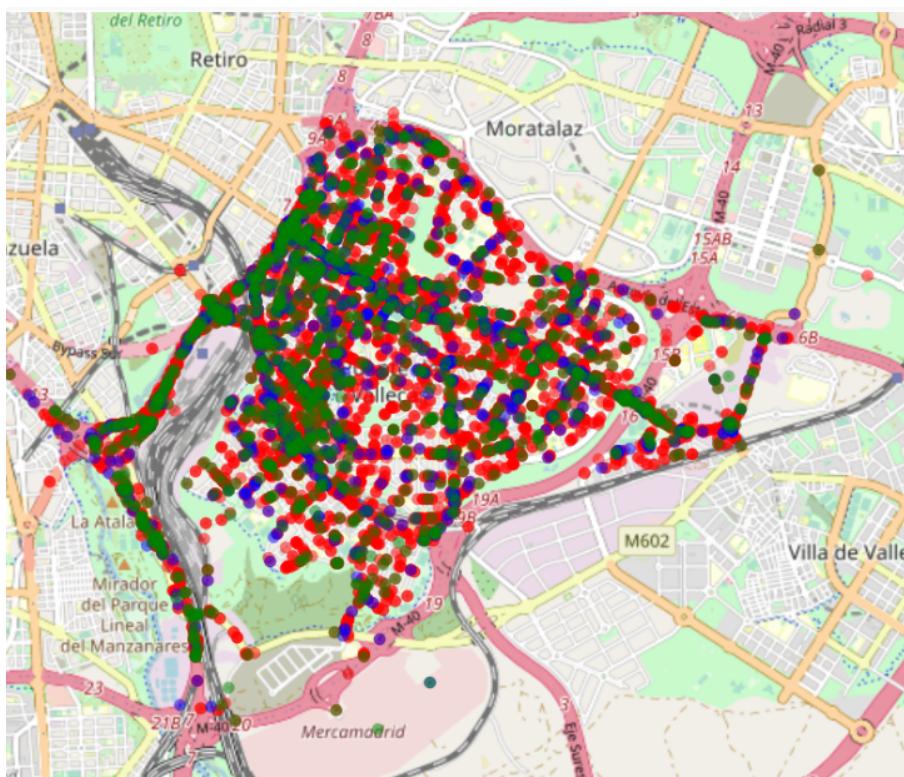
**FIGURA 10B:** Mapa de accidentes en el distrito de Salamanca: Turismos (rojo), Motocicletas (azul) y Vehículos de movilidad personal (verde)



**FIGURA 11A:** Gráfico de sectores de frecuencia de accidentes por tipo de vehículo en el distrito de Puente de Vallecas



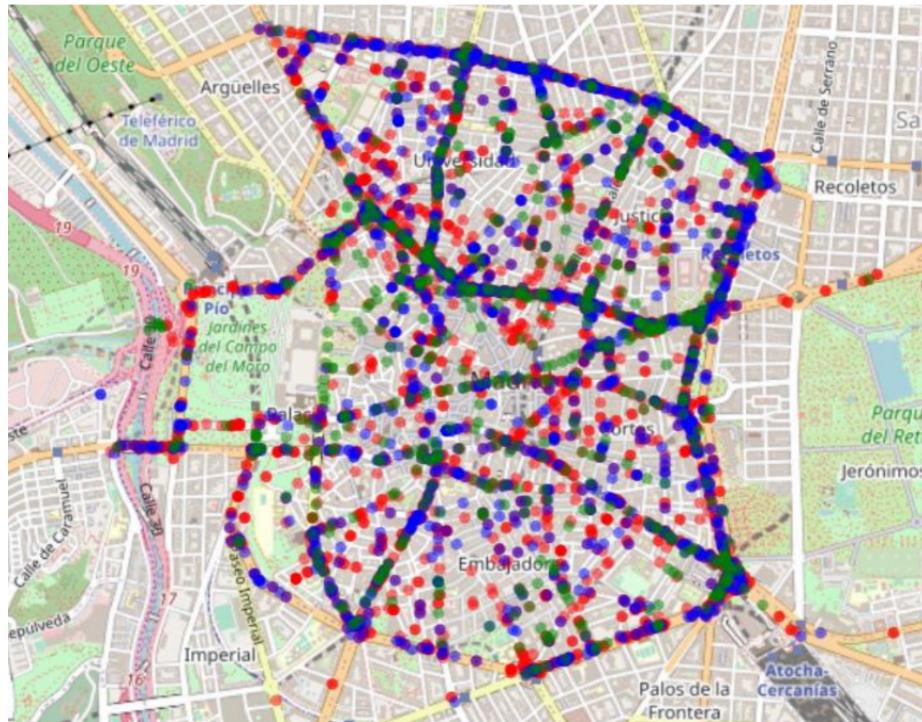
**FIGURA 11B:** Mapa de accidentes en el distrito de Puente de Vallecas: Turismos (rojo), Motocicletas (azul) y Camiones (verde)



**FIGURA 12A:** Gráfico de sectores de frecuencia de accidentes por tipo de vehículo en el distrito Centro

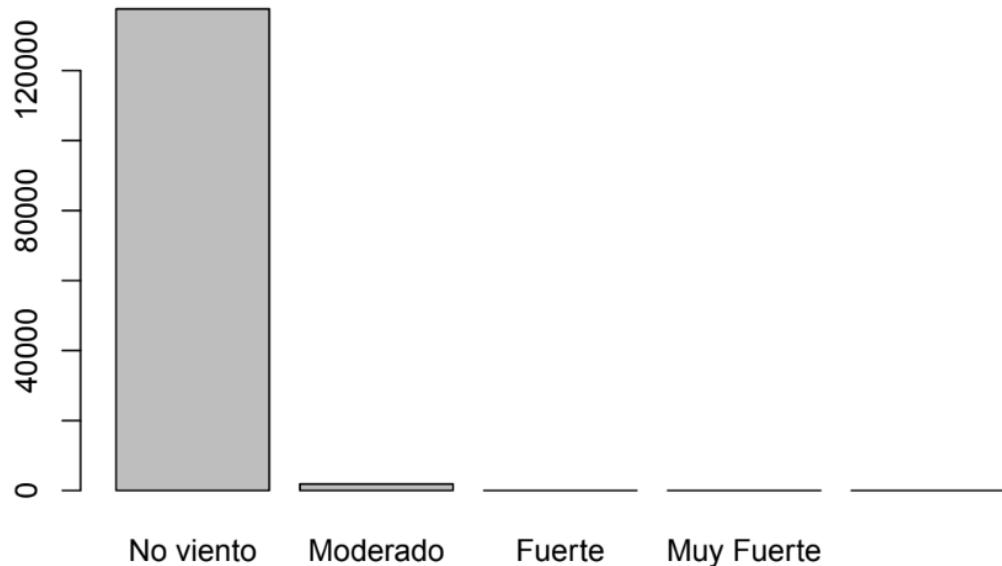


**FIGURA 12B:** Mapa de accidentes en el distrito Centro: Turismos (rojo), Motocicletas (azul) y Vehículos de movilidad personal (verde)

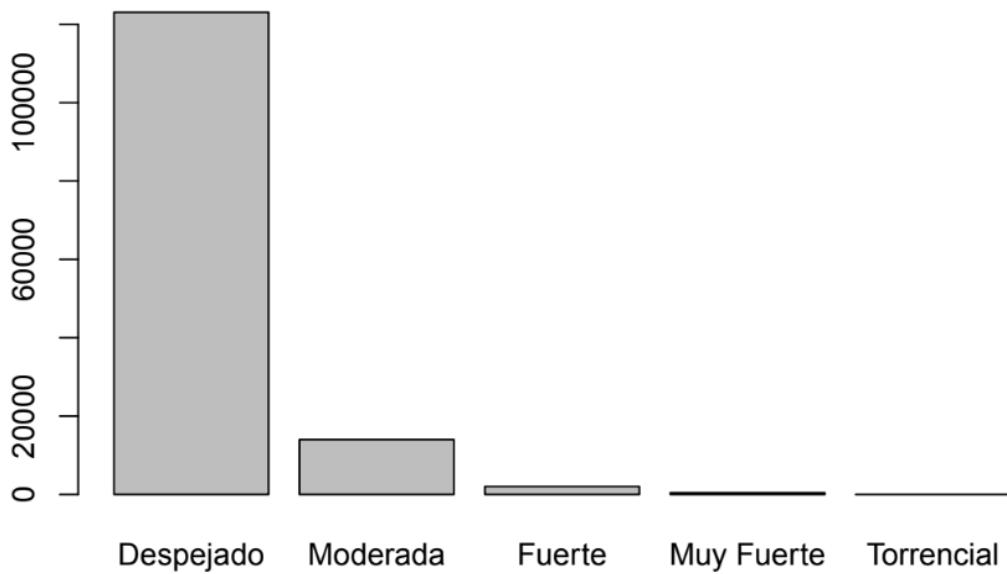


## Objetivo 3: Influencia de la meteorología

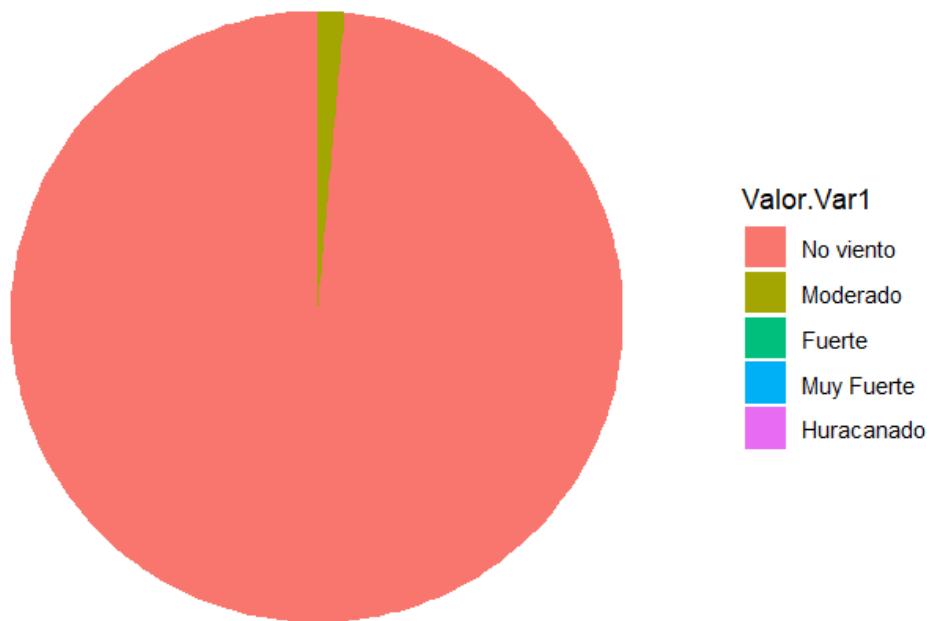
**FIGURA 13A:** Gráfico de frecuencia de accidentes según la situación del viento



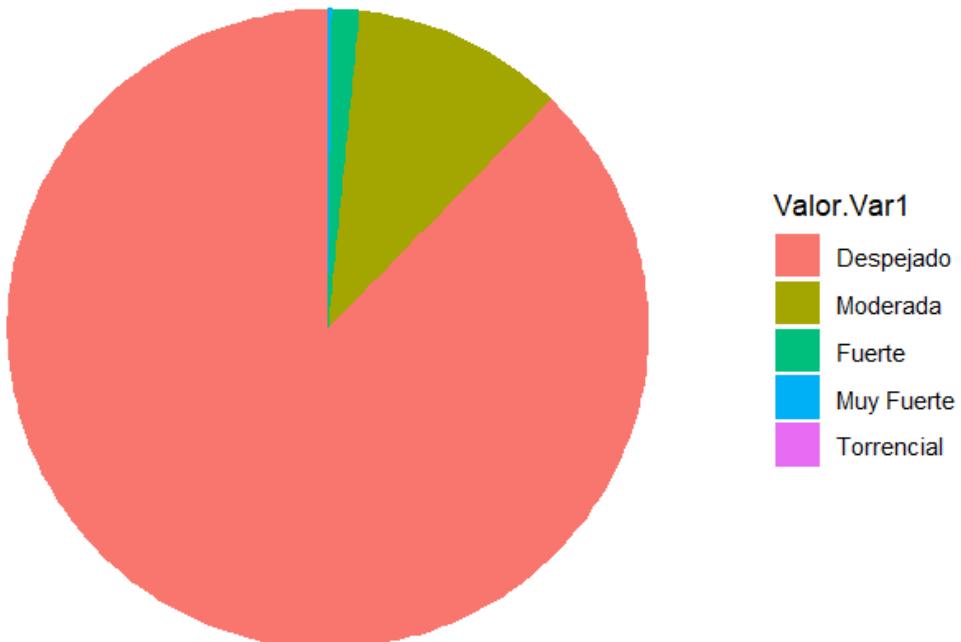
**FIGURA 13B:** Gráfico de frecuencia de accidentes según las precipitaciones



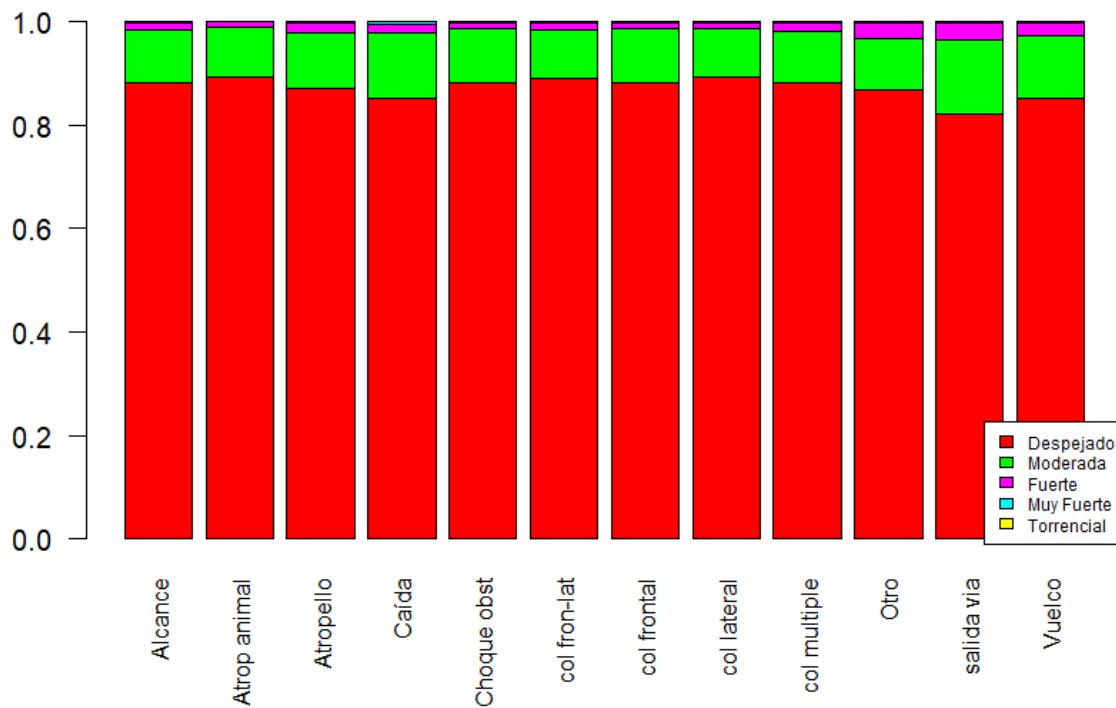
**FIGURA 14A:** Gráfico de sectores del porcentaje de días según la situación del viento



**FIGURA 14B:** Gráfico de sectores del porcentaje de días según las precipitaciones

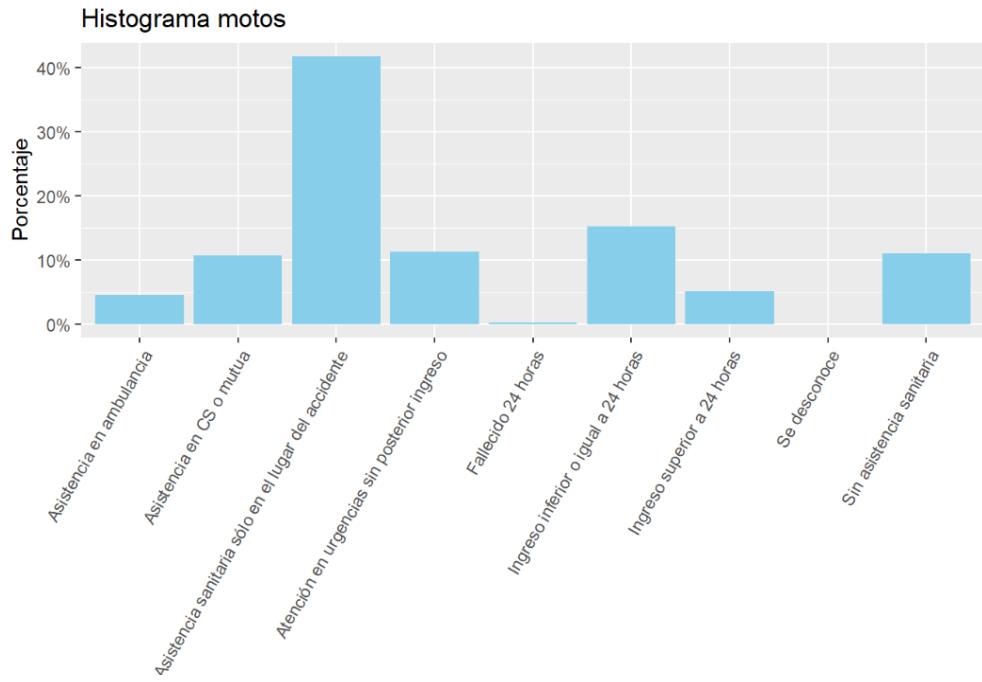


**FIGURA 15:** Frecuencia relativa de accidentes según el tipo de accidente, filtrado por nivel de precipitación

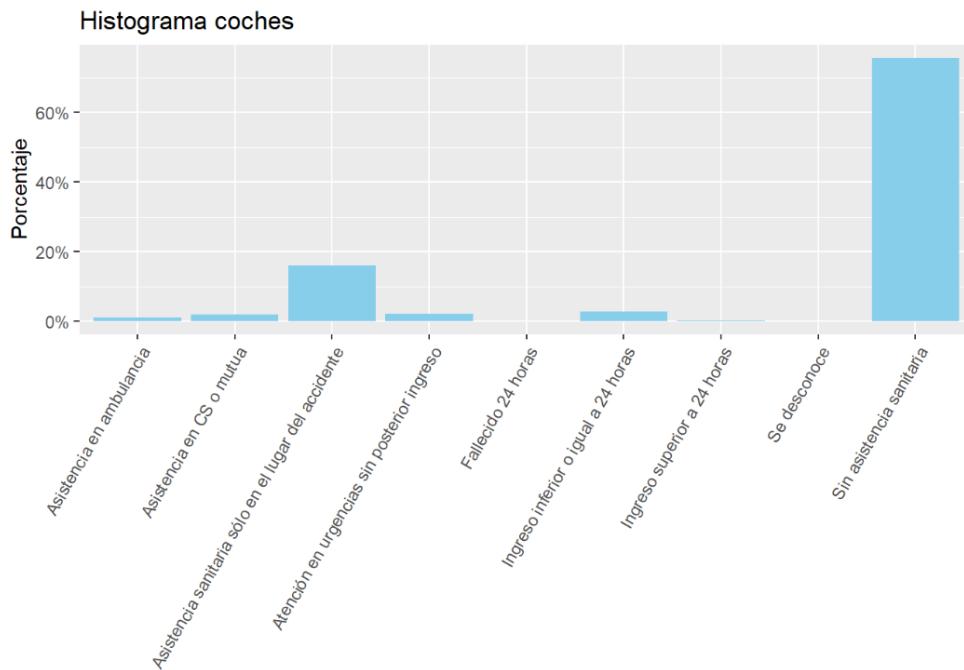


## Objetivo 4: Contraste turismo-moto

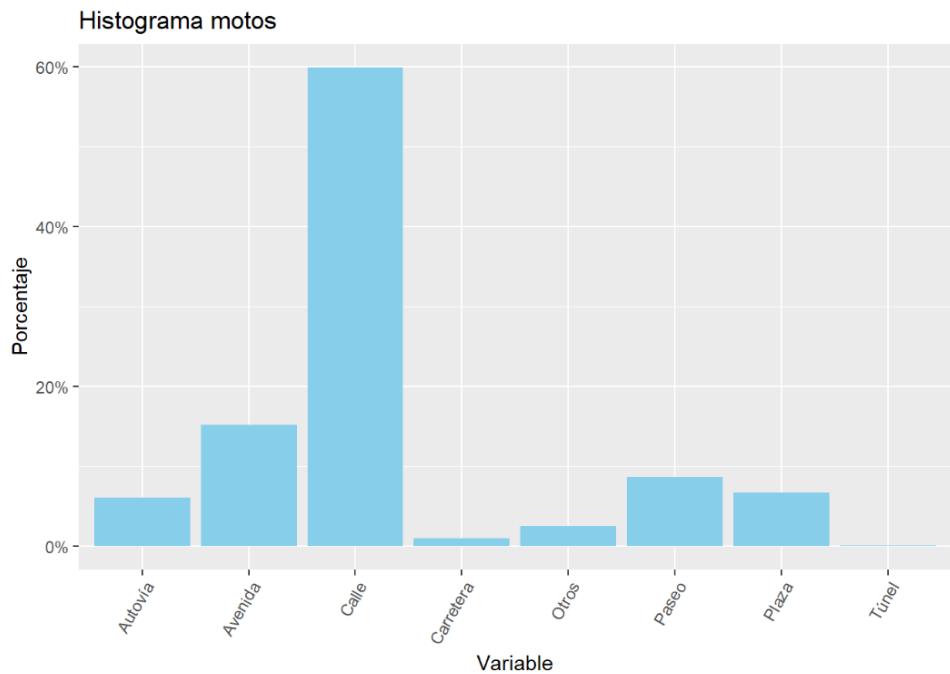
**FIGURA 16A:** Histograma de frecuencia relativa de accidentes de motocicletas según lesividad



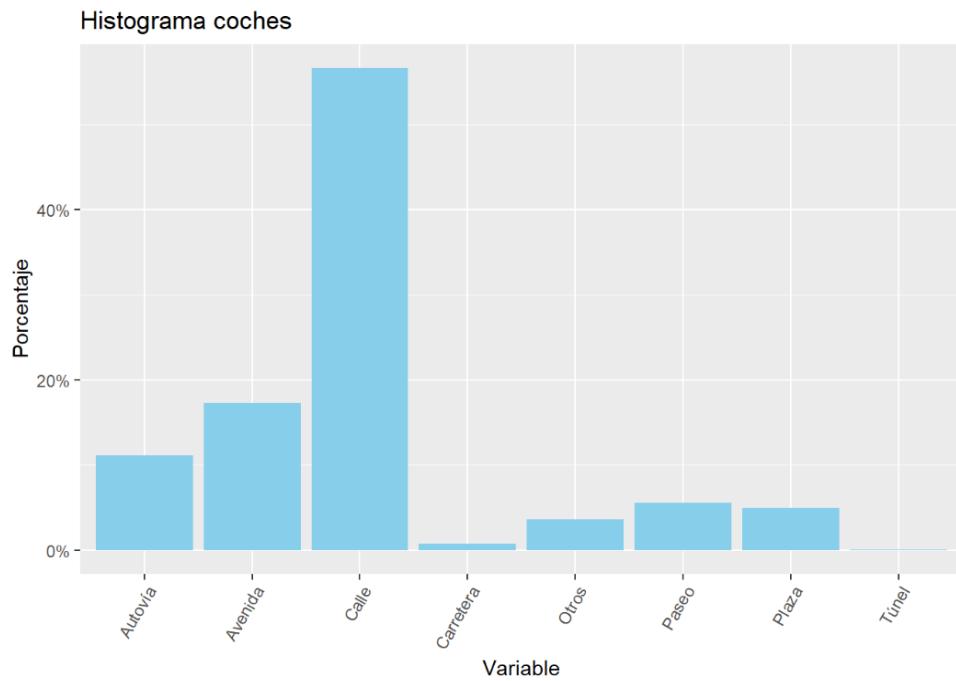
**FIGURA 16B:** Histograma de frecuencia relativa de accidentes de turismos según lesividad



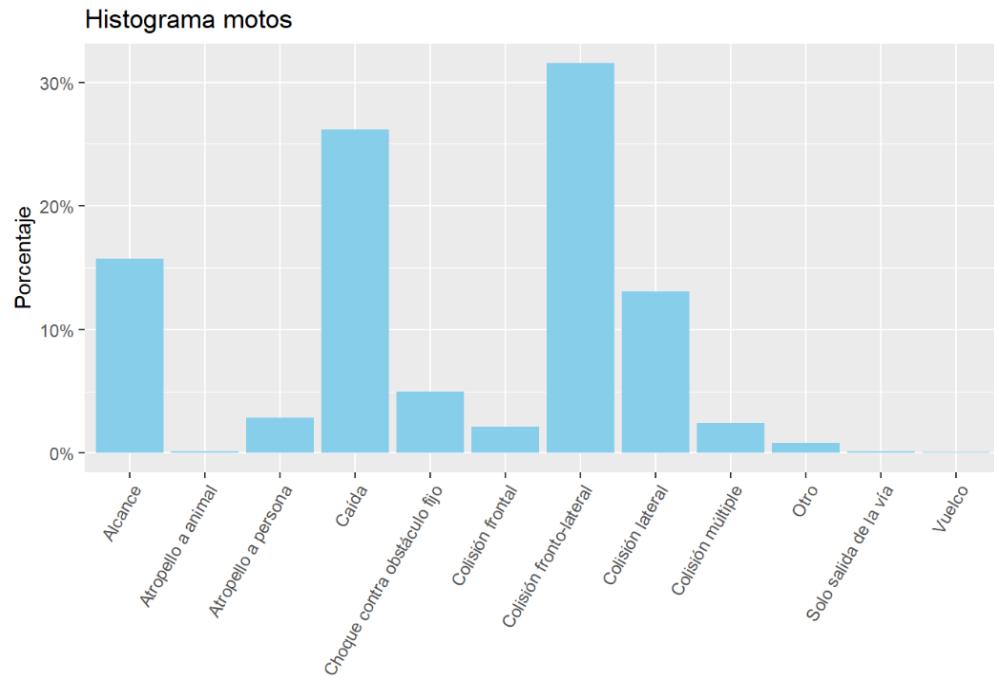
**FIGURA 17A:** Histograma de frecuencia relativa de accidentes de motocicletas según el tipo de vía



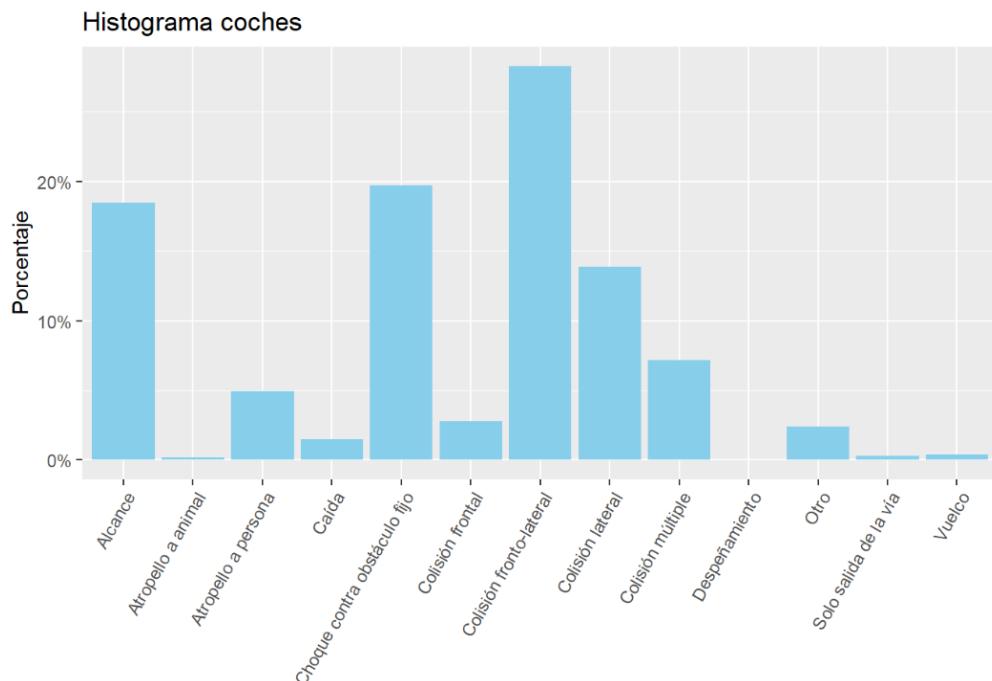
**FIGURA 17B:** Histograma de frecuencia relativa de accidentes de turismos según el tipo de vía



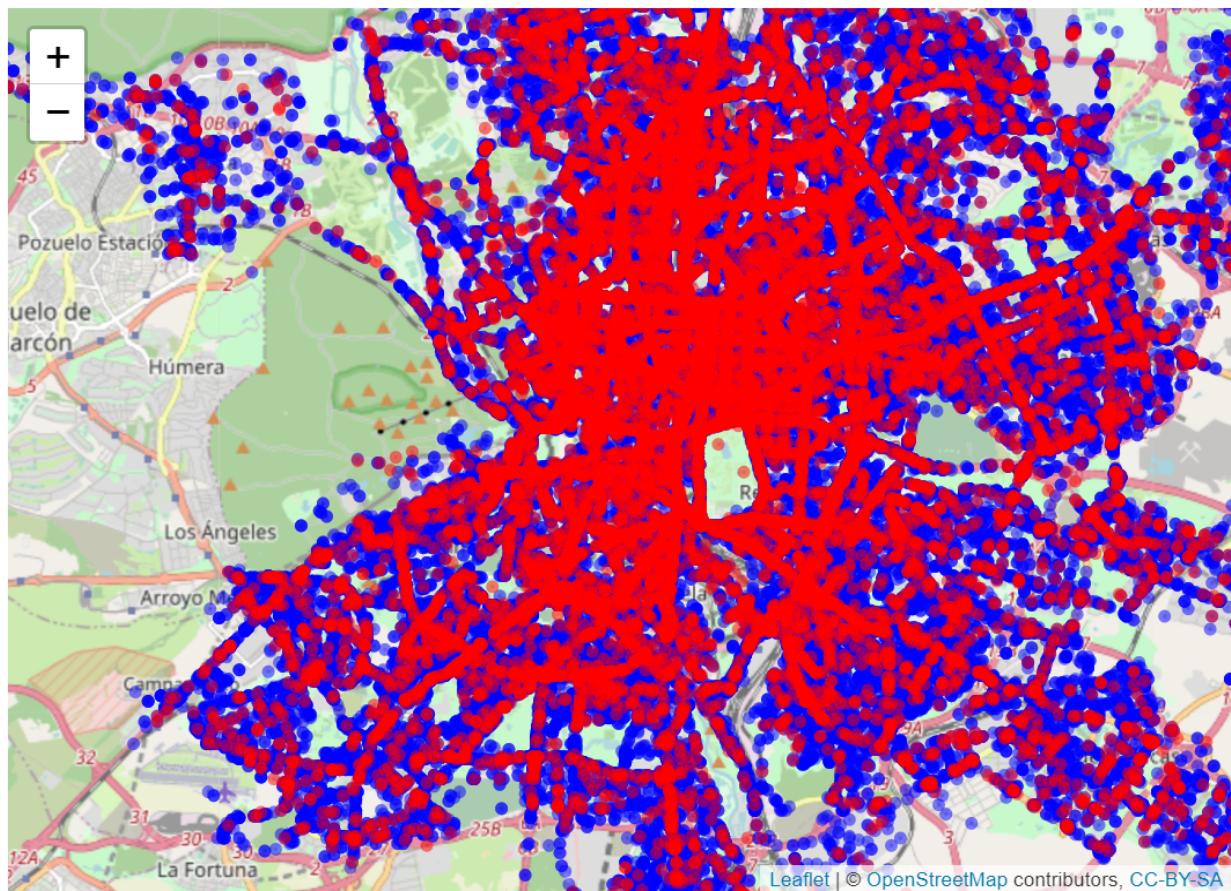
**FIGURA 18A:** Histograma de frecuencia relativa de accidentes de motocicletas según tipología del accidente



**FIGURA 18B:** Histograma de frecuencia relativa de accidentes de turismos según tipología del accidente

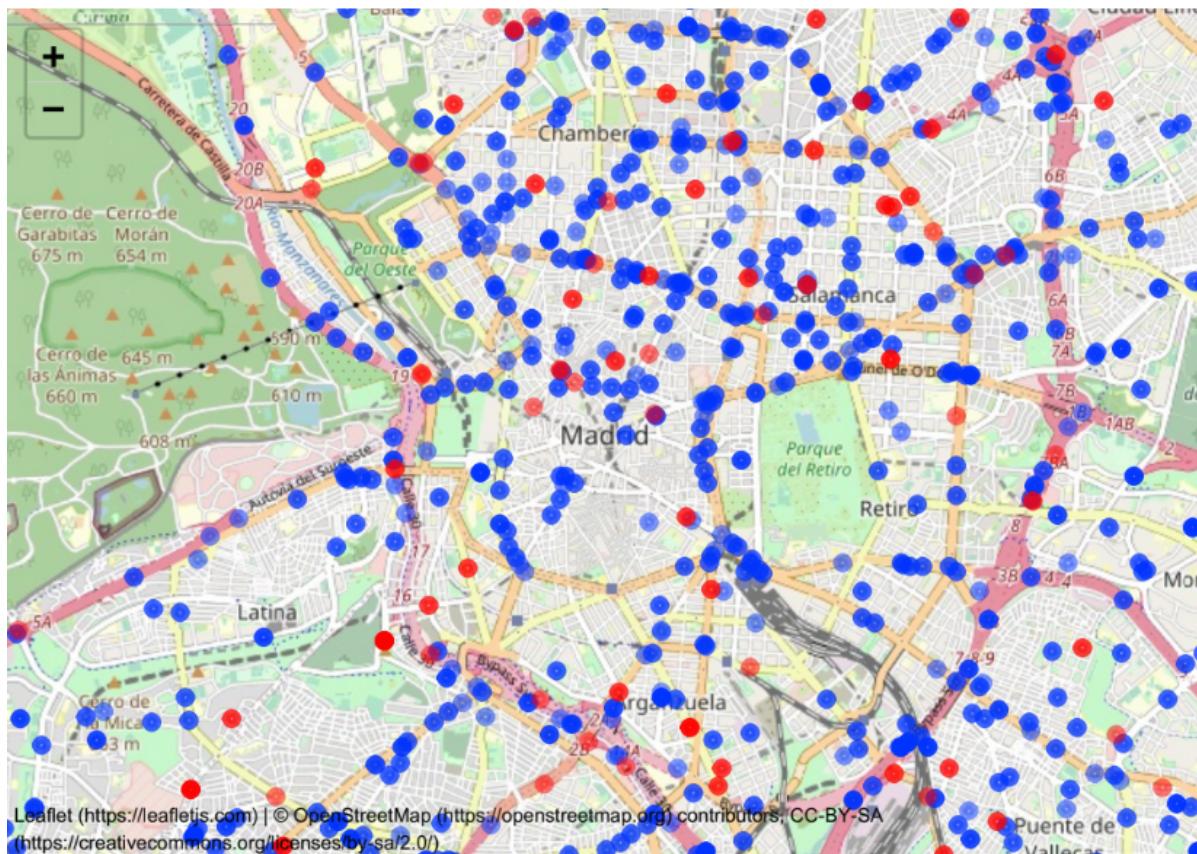


**FIGURA 19:** Mapa de accidentes en Madrid: Motocicletas (rojo) y Turismos (azul)



## Objetivo 5: Influencia del COVID-19

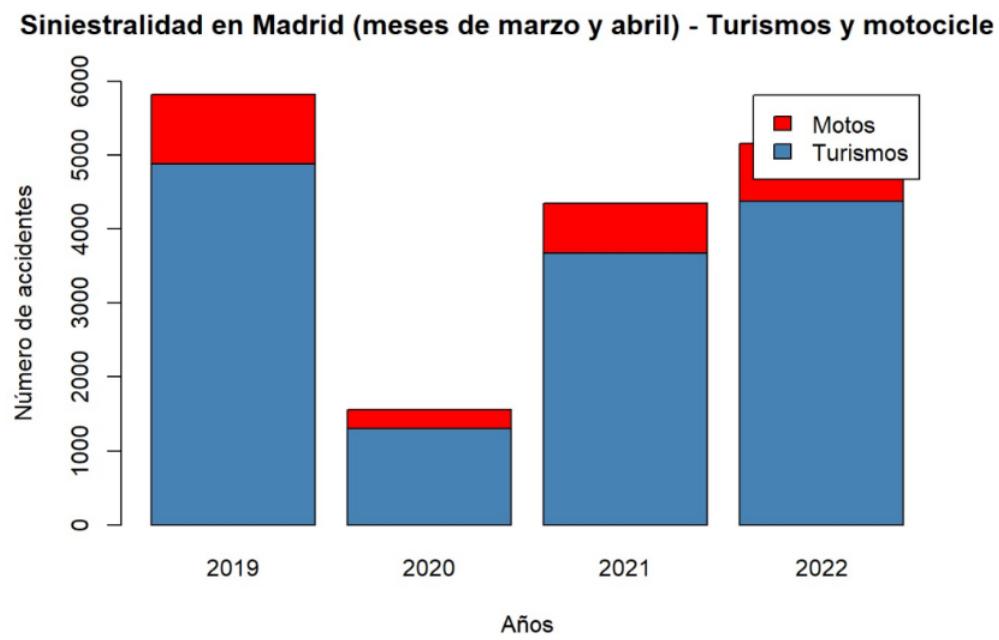
**FIGURA 20:** Mapa de accidentes en Madrid, filtrado por fecha: 16-30 de marzo de 2019 (azul) y 16-30 de marzo de 2020 (rojo)



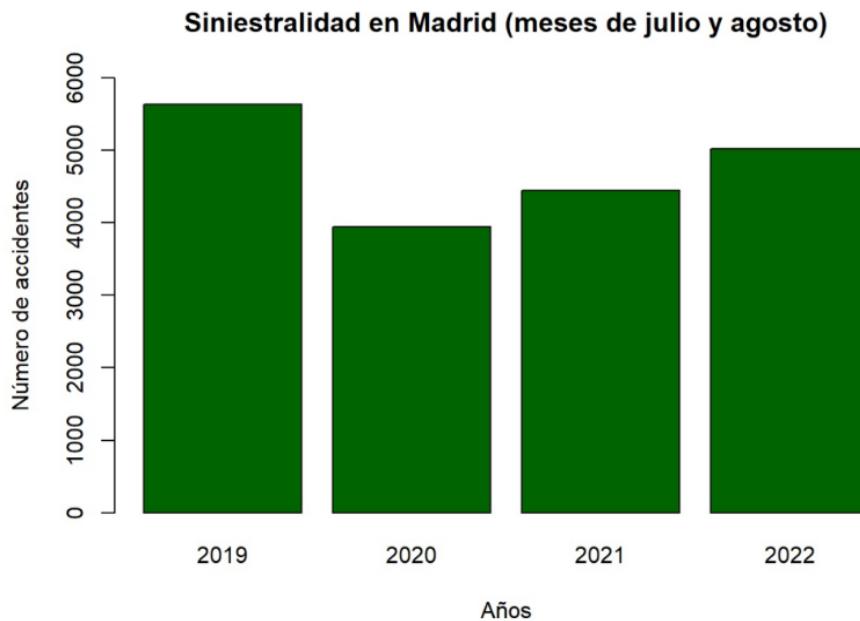
**FIGURA 21A:** Gráfico de frecuencia global de accidentes en marzo y abril



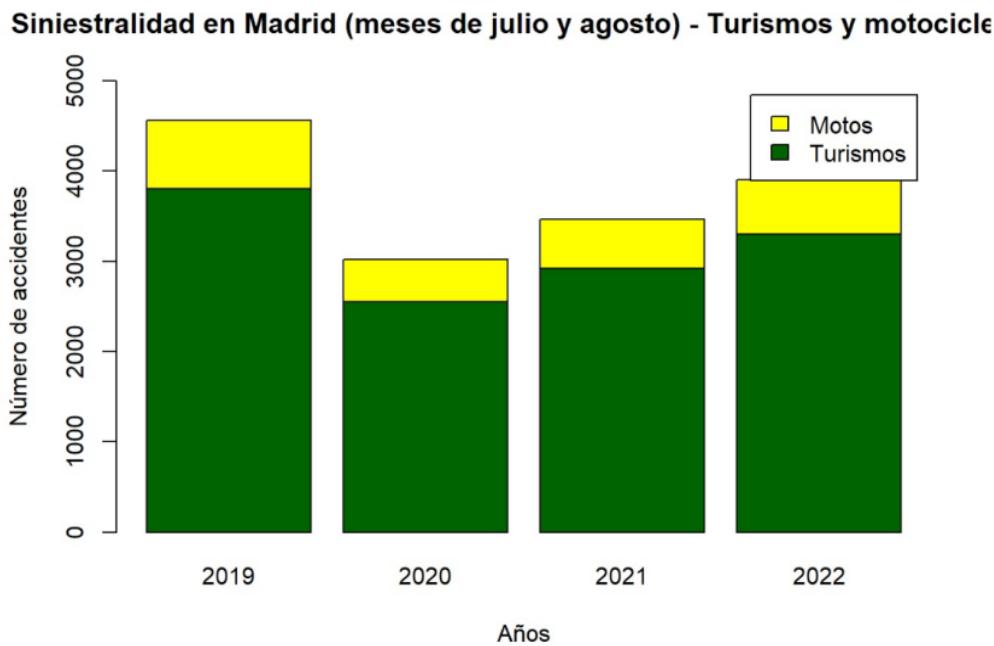
**FIGURA 21B:** Gráfico de frecuencia de accidentes en marzo y abril, filtrado por tipo de vehículo



**FIGURA 22A:** Gráfico de frecuencia global de accidentes en julio y agosto



**FIGURA 22B:** Gráfico de frecuencia de accidentes en julio y agosto, filtrado por tipo de vehículo



## PROYECTO II – Anexo 4

# Ficheros R Markdown



## Ciencia de Datos — Curso 2022/2023

Coral Montes, Adrián Rico, Juan Tomás, Marc Vicedo, Tingting Wu

---

A continuación adjuntamos los ficheros R Markdown de cada uno de los objetivos desarrollados en nuestro estudio. Debemos destacar que pueden existir ciertas diferencias de formato entre los ficheros, debido a que han sido ejecutados en ordenadores diferentes. Asimismo, algunos gráficos pueden presentar problemas tras su transformación a PDF.