EXAMEN DE TÉCNICAS PARA LA AGRUPACIÓN Y REDUCCIÓN DE LA DIMENSIÓN

Máster en Data Science, 2020-2021

Valentina Díaz Torres

20 DE ENERO DE 2021

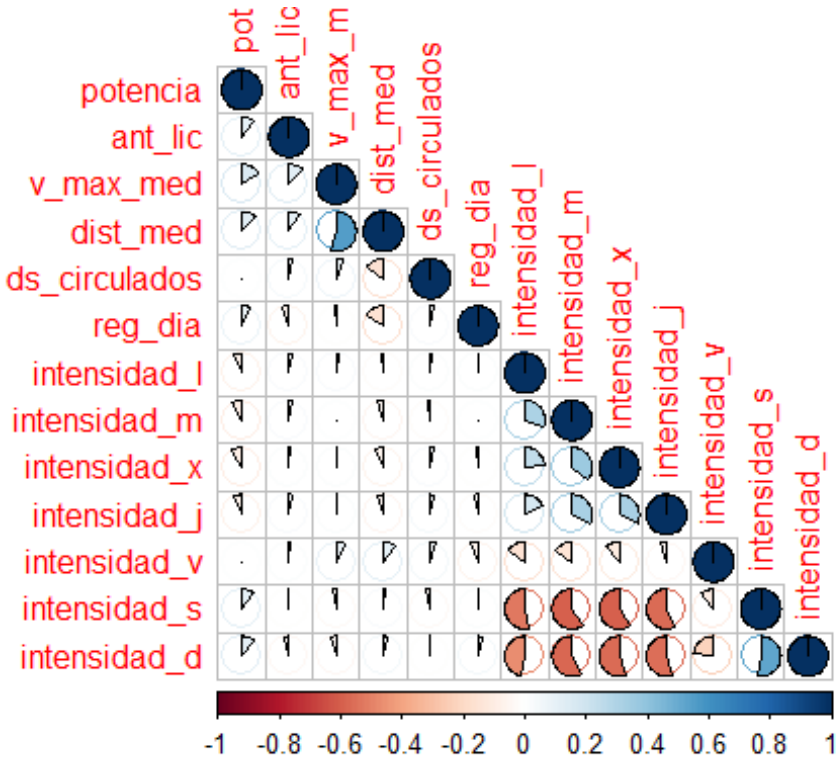
Objetivos e introducción.

La presente práctica práctica consiste en un análisis de 15 variables acerca de 20.000 conductores, que recogen información tal como la intensidad media de conducción, la potencia de los coches, velocidad máxima a la que conducen, entre otras.

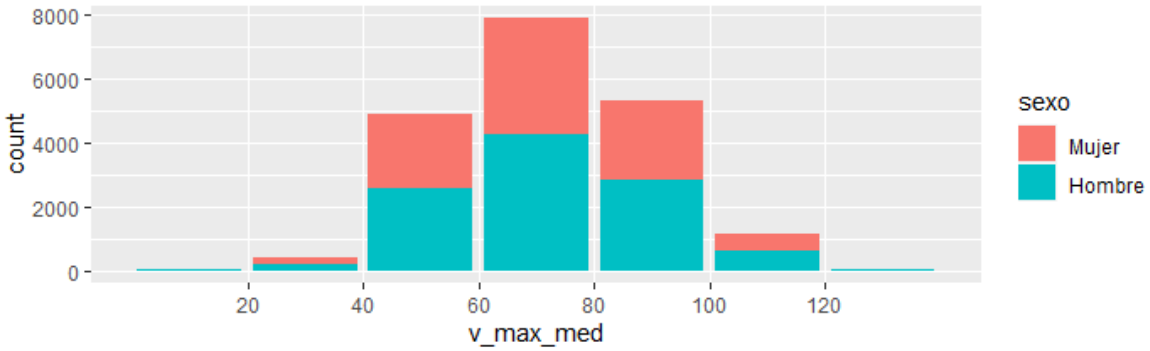
Se pretende, por un lado descubrir si existen posibles grupos o características comunes entre las variables uso diario del vehículo, recogidas bajo siete columnas del dataset que contienen información de los conductores sobre la intensidad del uso cada día de la semana. También, interesa estudiar si existe, por otro lado, relación entre las variables potencia, antigüedad de la licencia, velocidad máxima circulada, distancia recorrida y los registros diarios sobre el número de veces que mueven el coche los conductores. Por último, existen las variables sexo y siniestralidad, es decir, si ha existido o no algún siniestro. Respecto a estas dos últimas se pretende encontrar posibles intereses de cara al análisis objetivo.

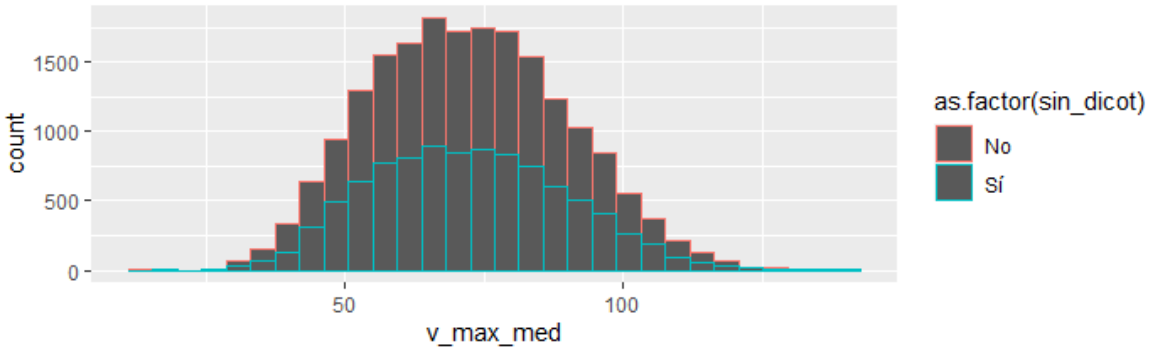
Para todo lo anterior, en primer lugar se realizará una reducción de la dimensión del dataset, seleccionando solo las 15 variables que interesan estudiar, de las 57 que contiene. Una vez reducido este,se realiza un análisis exploratorio con el fin de buscar anomalías en los datos y poder hacer limpieza y preparación de los mismos. Tras ello, se realiza el análisis cluster empleando métodos no jerárquicos, ya que son los más adecuados debido a las dimensiones del dataset y las necesidades computacionales de los algoritmos jerárquicos. Se han realizado los algoritmo K medias, PAM, CLARA y el de Segmentación borrosa (Fuzzy Clustering). No obstante, CLARA ha sido elegido como el algoritmo según el cual se van a extraer los datos para el análisis de las variables, quedando los otros tres como un apoyo para el mismo. Además, el número óptimo de grupos de conductores que se podían formar en cada uno de los análisis, tras varias pruebas ha sido de 2, en ambos casos. Es importante resaltar que las variables categóricas sexo y siniestralidad se han separado del análisis cluster para despúes compararlas con los resultados y buscar si realmente surge alguna relación importante, es decir si esta información puede aportarle un punto de vista interesante al análisis, mediante un análisis descriptivo.

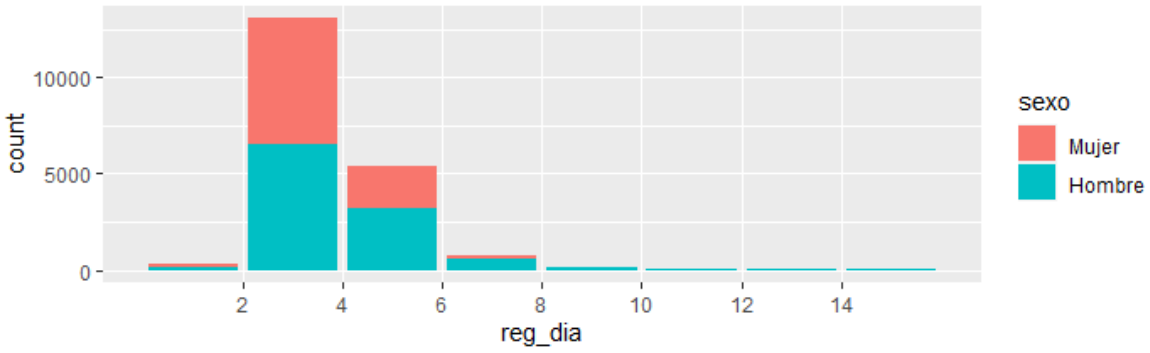
En primer lugar, se ha decidido empezar por una matriz de correlación, con el fin de observerar cúal es la relación entre las variables, de cara a futuros análisis. Se ha encontrado que la intensidad en los días entre semana tiene una correlación altamente negativa con los sábados y domingos. Estos se explicaría porque hay mucho menos tráfico los fin de semana, ya que, por lo general, menos personas van a trabajar en estos días. En cuanto a correlación positiva se podría destacar la de distancia media y velocidad media.

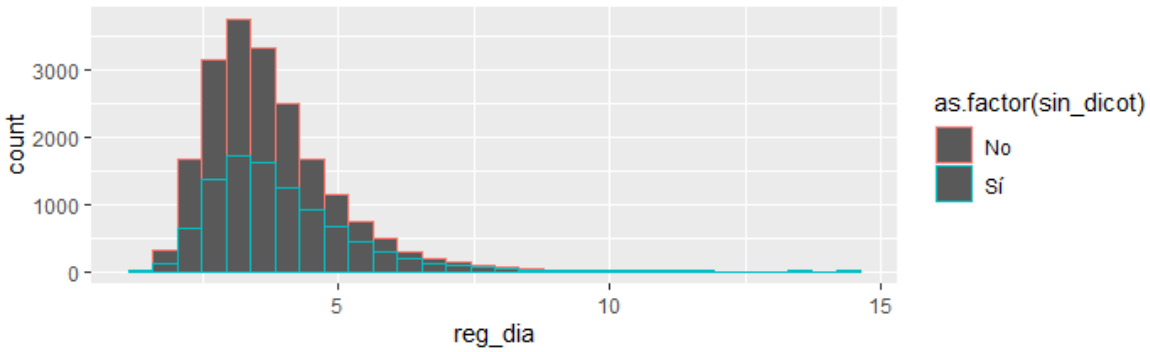


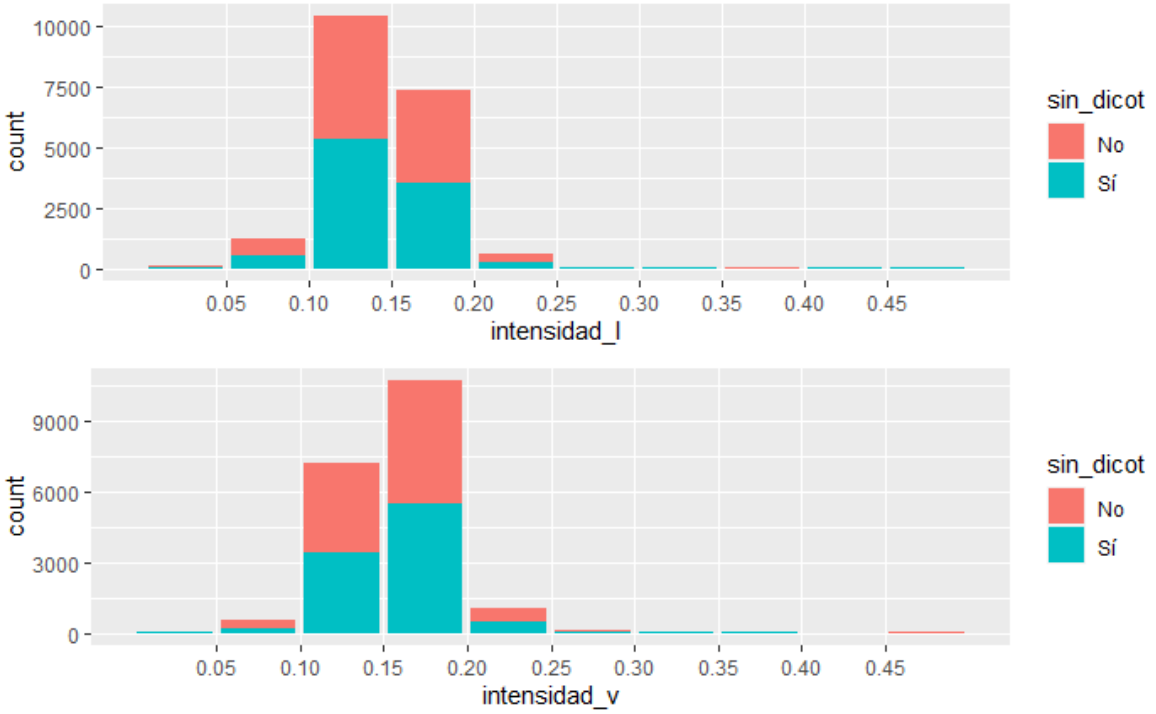
Tras ello, se realiza un análisis exploratorio, de modo gráfico, con el fin de, visualmente, observar la relación entre algunas variables pertinentes. En los siguientes gráficos se muestra la relación de siniestralidad y sexo con las variables velocidad, intensidad los lunes, intensidad los viernes y registros diarios. Los hallazgos encontrados son que, respecto al sexo, por lo general, se podría presuponer que las mujeres, en los datos estudiados, conducen a velocidad más moderada que los hombres, tienen potencia media en el coche y tienen más registros diarios entre 2 y 6 veces. Los hombres, son los únicos con registros a velocidad superior de 120 km/h, los únicos con potencia superior a 150 y de los que hay registros diarios de hasta 10 veces. Respecto a la intensidad según los días de la semana se ha pretendido comprobar si la intensidad era mayor finalizando o empezando la semana, que suele haber más tráfico y los siniestros que había estos días. Los resultados encontrados son bastante equitativos los dos días, aunque parece que los viernes hay más registros de mayor intensidad y más siniestralidad, aunque no es una diferencia muy significativa.


8000
6000
4000
2000
0
count
20
40
60
80
100
120
v_max_med
sexo
Mujer
Hombre

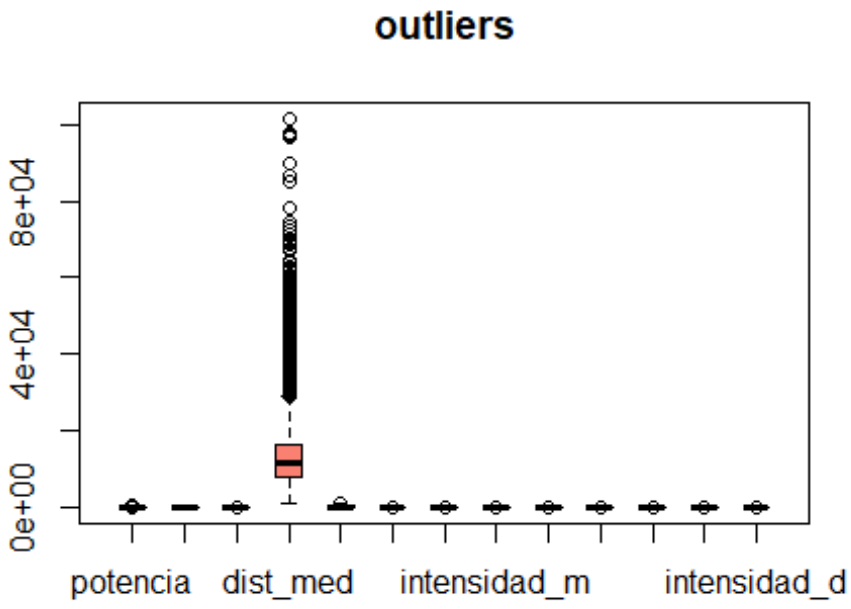

1500
1000
500
0
count
50
100
v_max_med
as.factor(sin_dicot)
No
Sí


10000
5000
0
count
2
4
6
8
10
12
14
reg_dia
sexo
Mujer
Hombre


3000
2000
1000
0
count
5
10
15
reg_dia
as.factor(sin_dicot)
No
Sí



Es necesario también, hacer una análisis de los valores atípicos que pudiesen existir en el dataset. Se observan que la variable dist_med contiene muchos. No obstante, tras varios análisis, se ha decidido no eliminarlos, ya que se podría estar perdiendo mucha información, en este caso sobre la distancia media que recorren los conductores, hay algunos que podrían hacer un trayecto más largo.



Debido que en los objetivos de esta práctica se busca estudiar un conjunto de variables separadas, por un lado todas las de intensidad, en los 7 días de la semana, y por otro, el resto, se van a dividir en dos subgrupos, para realizar análisis cluster por separados. “c1” compondría el primer grupo y “c2” el segundo.

Además, debido a que la base de datos está compuesta por cerca de 20.000 observaciones, en algunas casos puede resultar muy difícil calcular ciertas operaciones. Es por eso, que se han creado dos subgrupos más, compuestos por una muestra de 1000 observaciones del dataset.

Si las variables presentan fuertes variaciones de rango o una alta variabilidad,sería necesario escalarlas. En este caso, vemos como las distintas variables de intensidad ya vienen escaladas, mientras que otras como , días circulados o distancia media no. Es por esto, que solo se van a escalar las del primer grupo.

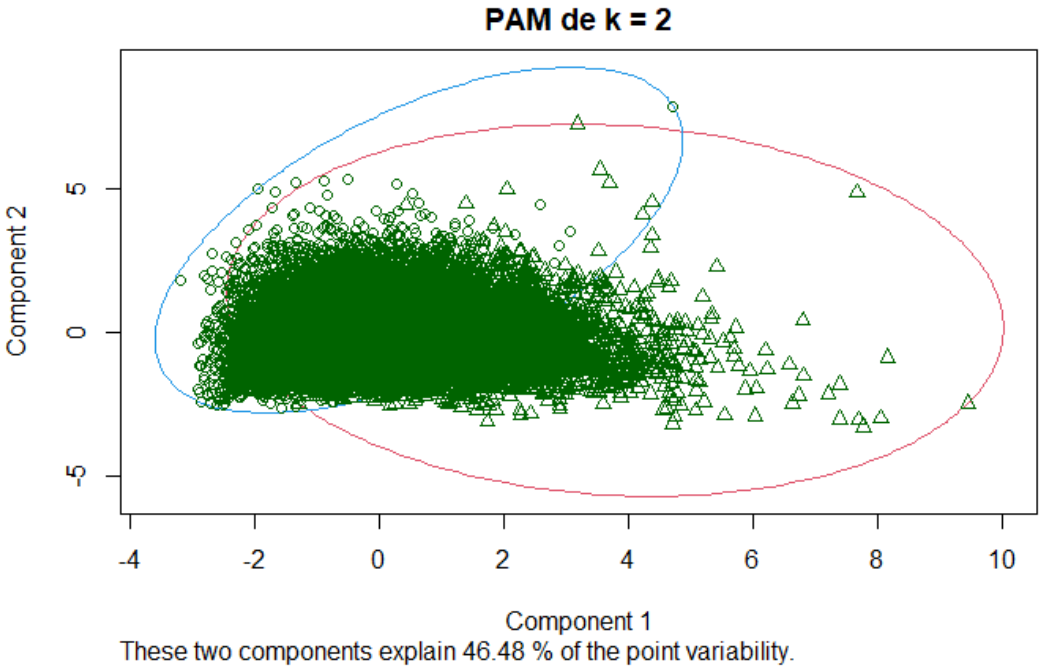
Algoritmos y Análisis Cluster

se realiza, mediante el estadístico Hopkins, una prueba para comprobar si tiene sentido o no hacer un análisis cluster y por tanto, si las variables se pueden agrupar entre ellas. El valor de este, debe ser lo más cercano a 0 para que el clustering tenga sentido, de ser próximo a 0.5, significaría que las variables son bastante homogéneas y que por tanto no se podrían agrupar. En ambos casos, el estadístico se encuentra muy por debajo del 0.5. No obstante, en el grupo de intensidad, este es menor (0.12 frente a 0.23 en el caso del primer grupo). Esto podría significar que entre estas variables de intensidad se pueden llevar a cabo dos grupos claramente más definidos, que en el caso de las otras variables.

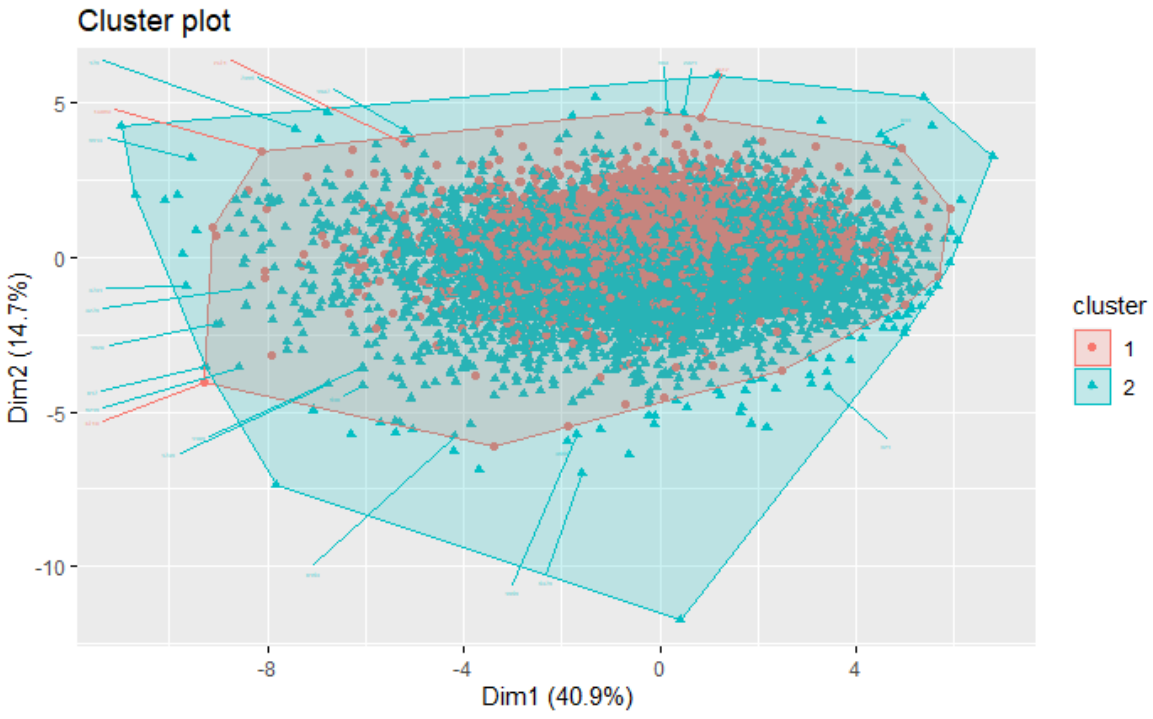
Como se ha dicho anteriormente, se han realizado cuatro algoritmos diferentes para analizar las variables, no obstante, debido a que unos son más adecuados que otros, se mostrarán solo dos de ellos. Por un lado, el algoritmo PAM (Partitioning Around Medoids), resuelve un problema principal del algoritmo k medias. La base de datos presente contiene un número importante de ouliers, como se ha decidido no modificarlos, es necesario elegir un algoritmo que corrija ese problema y que no sea muy sensible a ellos, como el k media lo es. El algoritmo PAM lo hace. Además, existe el algoritmo CLARA, que es muy similar a este anterior, pero trabaja con muestras, por lo que es perfecto para una base de datos de las dimensiones de la presente. Es por esto que este último será el elegido para hacer grupos entre las variables.

Algoritmo PAM (Partitioning Around Medoids)

Los dos clusters del primer grupo quedarían representados así:



Los dos clusters segundo grupo tendrían este aspecto:



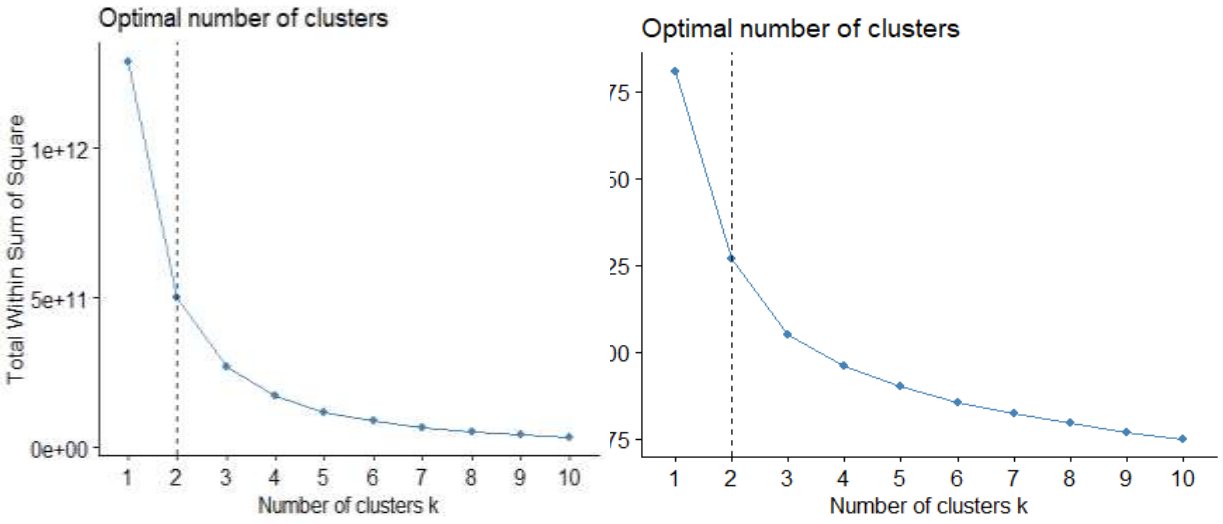
En ambos casos, los grupos aparecen bastante solapados, se necesitaría obserbar ambos gráficos desde otro plano para diferenciar los dos grupos. En el caso de las variables de intensidad, es decir, el Segundo gráfico, se observa cómo todos los puntos se encuentran muy juntos en el centro del mapa, muy mezclados, pero que el cluster dos es mucho más grande que el primero. También, la calidad de la representación en este ultimo es major que en el anterior.

Algoritmo CLARA (Clustering Large Applications)

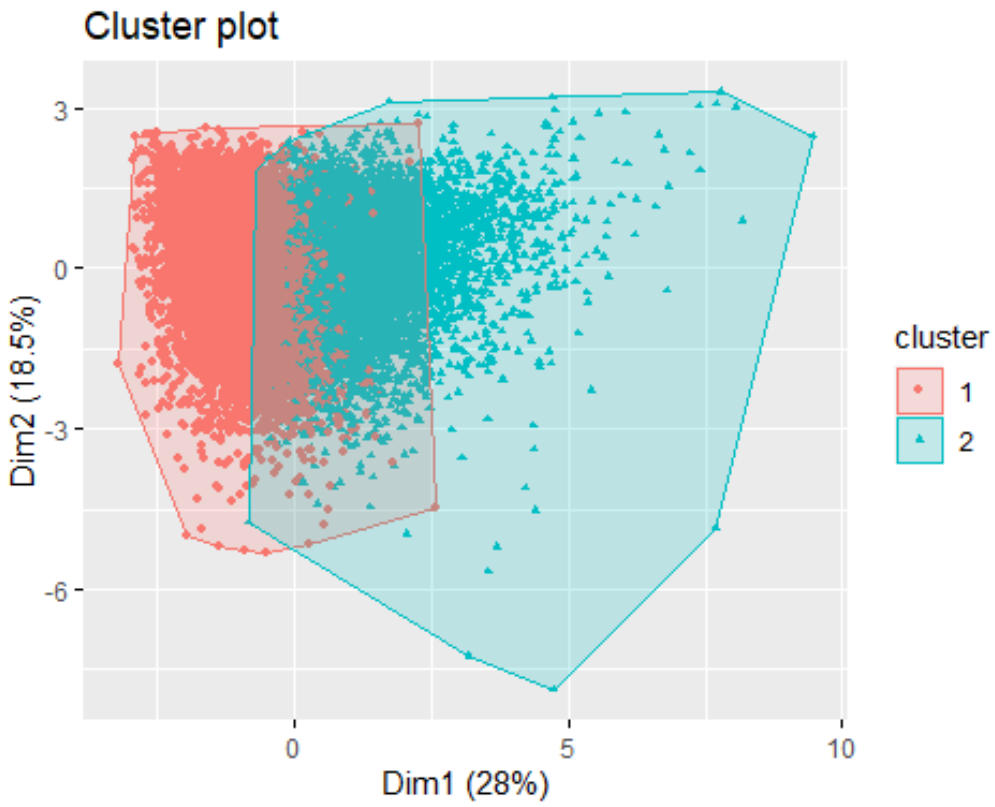
El algoritmo CLARAa ha sido elegido como idóneo para trabajar con la base de datos actual. Esta está compuesta por un gran número de observaciones, por lo que algoritmos computacionalmente intensos son muy difíciles de aplicar, dado los recursos limitados.

Para este algortimo se usa una muestra de la base de datos. Esta ha sido establecida como 1000 observaciones, ya que es el valor que computacionalmente se adapta a las capacidades y recursos presentes.

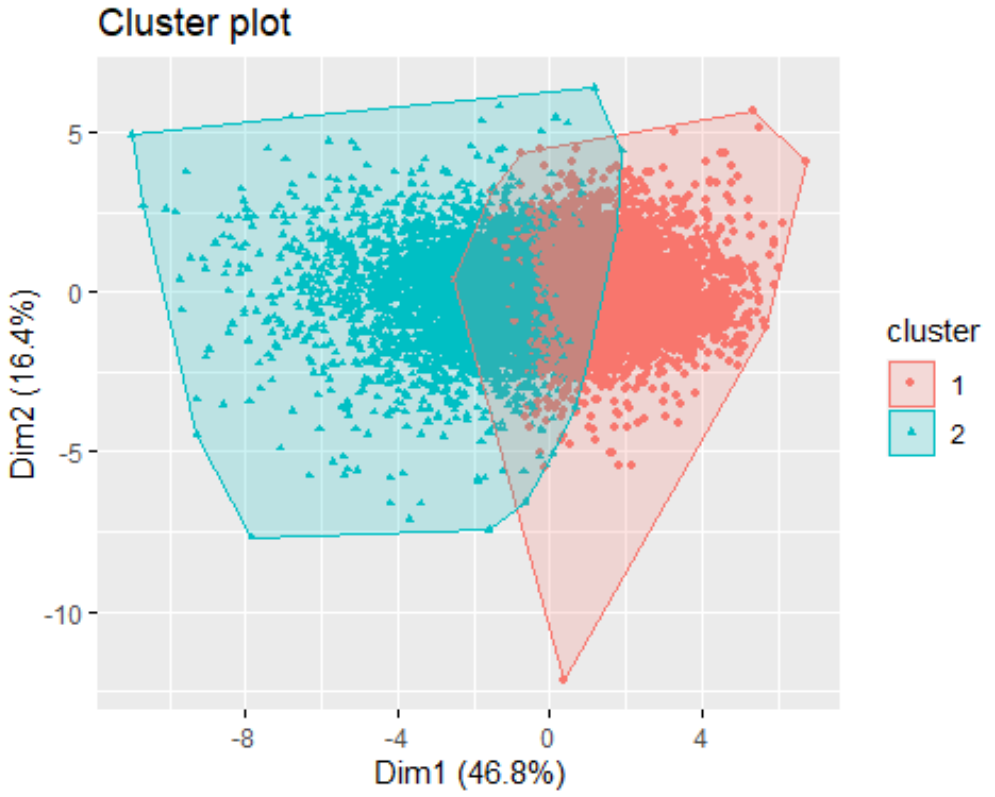
Antes de realizar el algoritmo final, clara, se calcula el número óptimo de clusters en cada grupo. Tras distintas pruebas, los resultados obtenidos han sido 2 en ambos casos, no obstante, quizá tres grupos también podrían existir según los siguientes gráficos.



En el análisis clara, se muestra la representación de los dos grupos en dos dimiensiones, donde los puntos se ubican según la distancia euclídea, anteriormente calculada y se encuadran en un grupo u otro. Del primer grupo la representación muestra que ambos grupos se solapan en cierto modo y que uno es mucho más grande que el otro. Además, la calidad de la representación no supera el 50%.



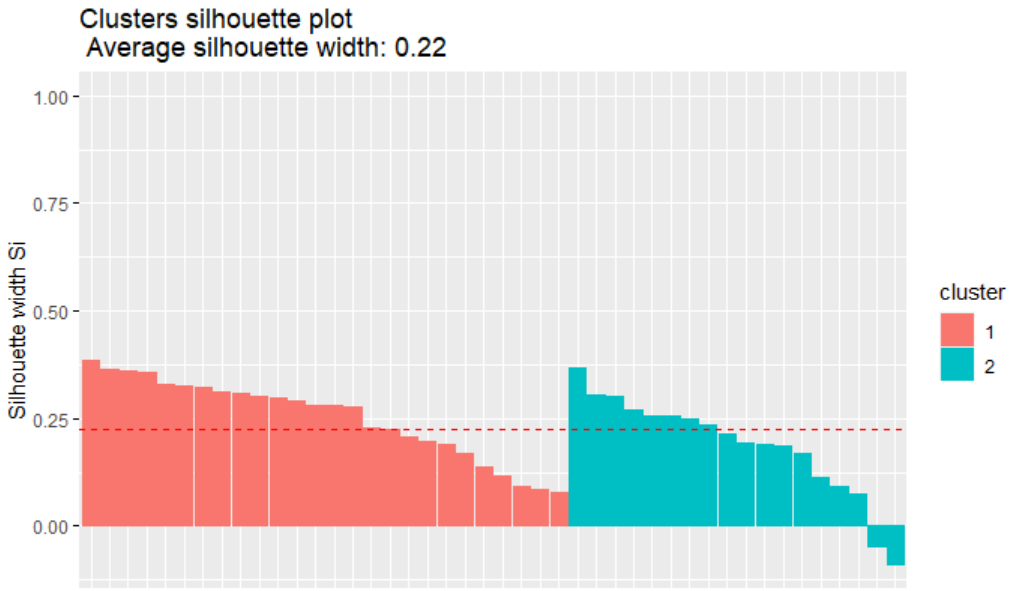
En el caso de las variables de intensidad, hay dos grupos claramente definidos y diferenciados, siendo estos, también, más homogéneos en cuanto a tamaño. La calidad de la representación es bastante mejor que en el caso anterior.



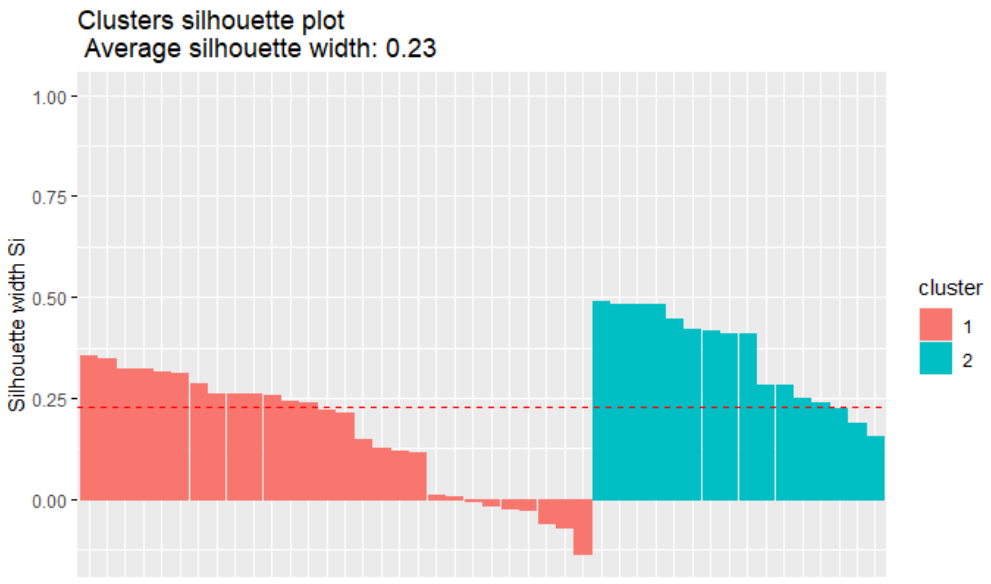
En este algoritmo, los grupos quedan major representados, se aprecia más cada uno de ellos, especialmente en las variables de intensidad. Este ha sido el algoritmo elegido para agrupar las observaciones de los conductors y extraer información de cada uno de estos grupos.

Además, se ha estudiado la silueta de ambos grupos para medir su similaridad. La anchura del gráfico mide la similaridad media de cada observación, así como la similaridad media del grupo.

La silueta del primer conjunto de variables es la siguiente:



La silueta del conjunto de variables relacionadas con la intensidad es:



Los valores cercanos a uno mostrarían una Buena clasificación y los cercanos a 0 significarían que son aquellas observaciones frontera entre un grupo y otro. Aquí se puede observar, en ambos gráficos, que no son muy tendentes a 1 y que están más cercanas a 0, lo que podría decir que son grupos formados por una clasificación no muy precisa. El major clasificado es el Segundo cluster del grupo de intensidad. También, en ambos casos hay valores negativos, que supondrían una clasificación errónea, sobre todo, en el caso del primer cluster de intensidad. No obstante, estos valores no han sido eliminados.

Conclusiones y hallazgos

Por último, se ponen en relación, a modo de conclusion, todas las variables y los grupos que se han formado, con el fin de analizar similaridades y conocer qué variables componen cada grupo y qué características tienen estos.

Respecto a las variables sexo y siniestralidad, además de lo ya comentado al inicio de la práctica, se ha comprobado que el número total de hombres estudiados es de 10473 y el de mujeres 9211. Además, hay 10057 casos que no han sido siniestro y 9627, por lo que los datos están bastante equilibrados y hay un alto caso de siniestros, en comparación con los no siniestros.

En el primer conjunto de variables, encontramos que, en el primer cluster se encuentran 12.246 conductores y en el segundo 7438. Respecto al cojunto de variables de intensidad, el primer grupo lo componen 11.653 conductores y el segundo 8.031, estando este segundo conjunto más equilibrado.

Además, en el primer grupo hay 5842 mujeres y 6404 hombres en el primer cluster y 3369 mujeres y 4069 en el segundo. En el grupo de variables de intensidad existen. En el grupo de intensidad 5824 mujeres y 5829 hombres en el primer cluster y 3387 muejres y 4644 hombres en el segundo. Entre las características, en el primer grupo, una vez que se han analizado ambos clusters, no se encuentran muchas diferencia entre hombres y mujeres.

En el caso de la siniestralidad, el análisis parece cobrar más sentido, ya que se denota cierta diferencia en los máximos y mínimos de cada variable, que indican que los conductores con siniestros, suelen hacer más recorrido diario, su distancia media es mayor, la velocidad y la potencia es algo más alta y circulan más días, lo cual tendría sentido, teniendo en cuenta que a mayor velocidad y más distancia recorrida, más probabilidad de un siniestro.

En conclusión, del grupo formado por las variables de intensidad, se ha podido extraer los hallazgos más relevantes del análisis, se han formado clusters más diferenciados y por tanto se ha visto como hay dos claros grupos, los conductores de entre semana, seguramente de trabajo y los de fin de semana. El grupo primero, está compuesto por carácterísticas diferentes, las cuales algunas han podido tomar relación con variables pero otras no. El sexo, por lo general no ha aportado mucha información, no hay una clara diferencia entre conductores hombres y mujeres, pero sin embargo la siniestralidad sí estaba más relacionada con las otras variables.