Universidad Del Valle de Guatemala


# ¿Cuáles son los factores que más influyen en la incidencia de accidentes en Guatemala?

Jonnathan Juárez, 15377
Sebastian Galindo, 15452
Eric Mendoza, 15002
José Ramírez, 15441

Curso de: MINERÍA DE DATOS.

Guatemala enero 28, 2019

# Índice

# Introducción

Dada la gran cantidad de personas que se movilizan en vehículos, y el aumento de vehículos que se integran al parque vehicular cada año, surge el problema de múltiples incidentes de tránsito. Es por esto que surge la idea de esta investigación que busca determinar los factores que se hacen presentes en los mismo. A continuación se presenta como parte introductoria la pregunta de investigación y justificación más detallada que dirigen este trabajo. Así, se incluye un análisis exploratorio de las variables, incluyendo estadísticas descriptivas de las variables numéricas relevantes, así como gráficas que permiten un mejor entendimiento de las variables categóricas. De estos se presentan hallazgos significativos de los datos. Finalmente, se muestra el análisi de grupos donde s verifica una forma adecuada de agrupar los datos.

# Descripción de situación problemática

Esta investigación se enfoca en el estudio y análisis de datos de la base de datos de accidentes de tránsito del Instituto Nacional de Estadística. El objetivo general de la investigación es hallar relaciones entre los hechos documentados para proveer hallazgos significativos que puedan ser útiles para la prevención de accidentes viales.

Nos encontramos en una situación en la que los accidentes viales ocurren a diario, y dado el incremento de carros que transitan cada vez hay más personas perjudicadas por dichos accidentes. Esta investigación ayudará a la prevención de accidentes de tránsito que ocurren. Los resultados serán de gran utilidad para la creación de nuevas campañas para la prevención de accidentes viales. Además, ayudará a futuras investigaciones que quieran hacer uso de la información obtenida en la investigación.

Los resultados recolectados pueden llegar para tratar situaciones similares en distintos departamentos de Guatemala. También, la investigación podría servir para replicar este tipo de análisis en países con

condiciones viales similares. Pero a pesar de ello es de importancia recalcar que los instrumentos para el análisis de datos a utilizar serán lenguajes de programación formales.

# Problema científico

En la actualidad, el departamento de Guatemala es un departamento que sufre de una sobrepoblación vehícular. Hoy en día, es normal ver que en una misma familia existan 2 o más vehículos. Esto induce que naturalmente exista un incremento en la carga vehícular y que, en particular, sea más notable en las famosas “horas pico”.

Debido a este notable incremento en la cantidad de los vehículos, la cantidad de accidentes de tránsito ha tenido un igual incremento. En una entrevista realizada a un ajustador de seguros para vehículos indicó: “A los 10 minutos de iniciada la hora pico, empezamos a recibir llamadas sobre colisiones.”. Por lo tanto, al ser cada vez más frecuentes los accidentes en Guatemala nos hallamos ante la necesidad de saber qué es lo que generalmente ocasiona que ocurran todos estos accidentes. Al ser tantos los posibles factores que influyen en un accidente (como el clima, estado del conductor, género del conductor, estado del vehículo, tipo de vehículo, etc..) necesitamos poder filtrar y encontrar los factores que poseen más influencia.

Actualmente no existen estudios o modelos que nos indiquen cuáles son las causas más comunes para los accidentes de tránsito. Al no contar con estos datos nos encontramos ante una situación en la que no podemos decidir qué medidas tomar para evitar dichos accidentes. Encontrar los factores que más influyen en la incidencia de accidentes de tránsito es una incógnita que de ser resuelta nos permitiría elaborar modelos que nos ayuden a evaluar distintos casos y escenarios, y tomar acciones preventivas sobre las mismas.

## Objetivo General

- Determinar cuáles son las variables/factores que más influyen en la incidencia de accidentes de tránsito.

### Objetivos Específicos

- Elaborar un modelo que describa la probabilidad de que un accidente ocurra dadas ciertas condiciones.
- Analizar y describir las tendencias existentes en los accidentes ocurridos en el año 2009 al 2017.

# Dataset

Se utilizó la base de datos del INE, de accidentes de tránsito, la cual contaba con los datos de 2009 a 2017 de los vehículos involucrados, fallecidos y hechos de tránsito. Para el análisis no fue necesario realizar una limpieza, ya que el dataset, había sido procesado previamente. A continuación se describen las variables de cada conjunto de datos.

*Vehículos involucrados*
Se observa en el Cuadro 1 que se tienen 18 variables relacionadas con los vehículos implicados en los accidentes registrados. Se cuenta con 54960 registros en este data set.

*Personas involucradas*
Se observa en el Cuadro 2 que se tienen 18 variables relacionadas con las personas involucradas con los accidentes registrados. Se cuenta con 74449 registros en este data set.

*Hechos de tránsito*
Se observa en el Cuadro 3 que se tienen 25 variables relacionadas con los accidentes registrados. Se cuenta con 45230 registros en este data set.

*Cuadro 1: Nombres de las variables del dataset de vehículos según su descripción*

| *Nombre de variable* | *Descripción* | *Tipo* |
|---|---|---|
| dia_ocu | Día ocurrido | Cuantitativa/Discreta |
| mes_ocu | Mes ocurrido | Cualitativa/Nominal |
| dia_sem_ocu | Día de la semana ocurrido | Cualitativa/Nominal |
| hora_ocu | Hora ocurrido | Cuantitativa/Discreta |
| depto_ocu | Departamento donde ocurrió | Cualitativa/Nominal |
| area_geo_ocu | Rural o urbana | Cualitativa/Nominal |
| zona_ocu | Zona del hecho | Cualitativa/Nominal |
| sexo_per | Sexo del fallecido | Cualitativa/Nominal |
| edad_per | Edad de la persoan | Cuantativa/Discretas |
| estado_per | Estado en que se encontraba en conductor | Cuantativa/Discretas |
| tipo_veh | Tipo de vehículo | Cualitativa/Nominal |
| color_veh | color del vehiculo | |
| modelo_veh | Modelo del vehículo | Cuantitativa/Discreta |
| tipo_eve | Tipo de evento del accidente | Cualitativa/Nominal |
| mayor_menor | Si el conductor es menor o menor | Cualitativa/Nominal |
| num_hecho | Número de hecho sucedido | Cuantitativa/Discreta |
| marca_veh | Marca del vehículo | Cualitativa/Nominal |
| g_edad_pil | No determinado | - |

*Cuadro 2: Nombres de las variables del dataset de personas implicadas en accidente según su descripción*

| *Nombre de variable* | *Descripción* | *Tipo* |
|---|---|---|
| num_hecho | Número de hecho sucedido | Cuantitativa/Discreta |
| día | Día ocurrido | Cuantitativa/Discreta |
| mes | Mes ocurrido | Cualitativa/Nominal |
| ano | Año ocurrido | Cuantitativa/Discreta |
| hora | Hora ocurrido | Cuantitativa/Discreta |
| dia_sem | Día de la semana ocurrido | Cualitativa/Nominal |
| edad | Edad del fallecido | Cuantitativa/Discreta |
| sexo | Sexo del fallecido | Cualitativa/Nominal |
| dept | Departamento del hecho | Cualitativa/Nominal |
| muni | Municipio del hecho | Cualitativa/Nominal |
| zona | Zona del hecho | Cualitativa/Nominal |
| area | Área rural o urbana | Cualitativa/Nominal |
| tipo_veh | Tipo de vehículo | Cualitativa/Nominal |
| marca_veh | Marca del vehículo | Cualitativa/Nominal |
| color_veh | Color del vehículo | Cualitativa/Nominal |
| modelo_veh | Modelo del vehículo | Cuantitativa/Discreta |
| tipo_eve | Tipo de evento | Cualitativa/Nominal |
| fall_les | Si el implicado falleció o solo se lesionó | Cualitativa/Nominal |

*Cuadro 3: Nombres de las variables del dataset de hechos de según su descripción*

| *Nombre de variable* | *Descripción* | *Tipo* |
|---|---|---|
| num_hecho | Número de hecho sucedido | Cuantitativa/Discreta |
| día_ocu | Día ocurrido | Cuantitativa/Discreta |
| mes_ocu | Mes ocurrido | Cualitativa/Nominal |
| ano_ocu | Año ocurrido | Cuantitativa/Discreta |
| hora_ocu | Hora ocurrido | Cuantitativa/Discreta |

| | | |
|---|---|---|
| dia_sem | Día de la semana ocurrido | Cualitativa/Nominal |
| edad_pil | Edad del fallecido | Cuantitativa/Discreta |
| sexo_pil | Sexo del fallecido | Cualitativa/Nominal |
| dept_ocu | Departamento del hecho | Cualitativa/Nominal |
| mupio_ocu | Municipio del hecho | Cualitativa/Nominal |
| zona_ocu | Zona del hecho | Cualitativa/Nominal |
| areag_ocu | Área rural o urbana | Cualitativa/Nominal |
| tipo_veh | Tipo de vehículo | Cualitativa/Nominal |
| marca_veh | Marca del vehículo | Cualitativa/Nominal |
| color_veh | Color del vehículo | Cualitativa/Nominal |
| modelo_veh | Modelo del vehículo | Cuantitativa/Discreta |
| causa_acc | El motivo del accidente | Cualitativa/Nominal |
| g_hora | Grupo de hora | Cualitativa/Nominal |
| g_modelo_veh | grupo modelo de vehículo | Cualitativa/Nominal |
| g_hora_5 | Grupo de hora (Intervalo de 5) | Cualitativa/Nominal |
| mayor_menor | si el inplicaod es mayor de edad | Cualitativa/Nominal |
| estado_pil | Estado del piloto | Cualitativa/Nominal |
| corre_base | base de correlativo | Cuantitativa/discreta |
| edad_quinquenales | edad en quinquenales | Cualitativa/Nominal |
| g_edad | Grupo de edad | Cualitativa/Nominal |

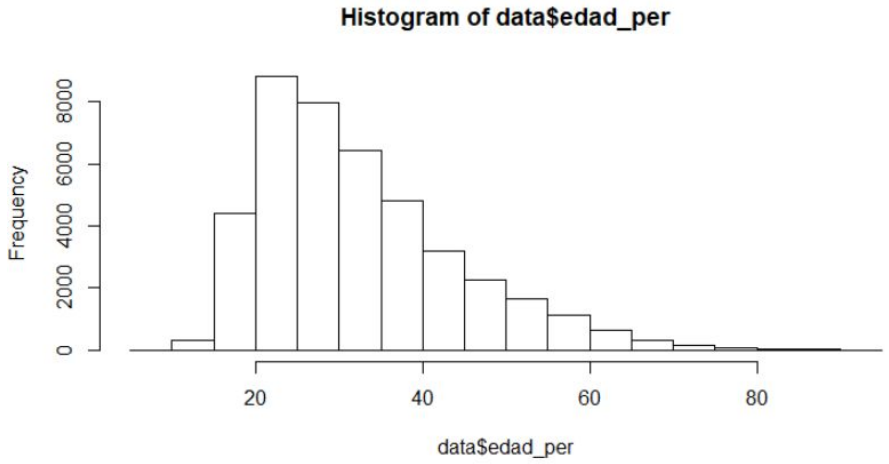
## Exploración de datos

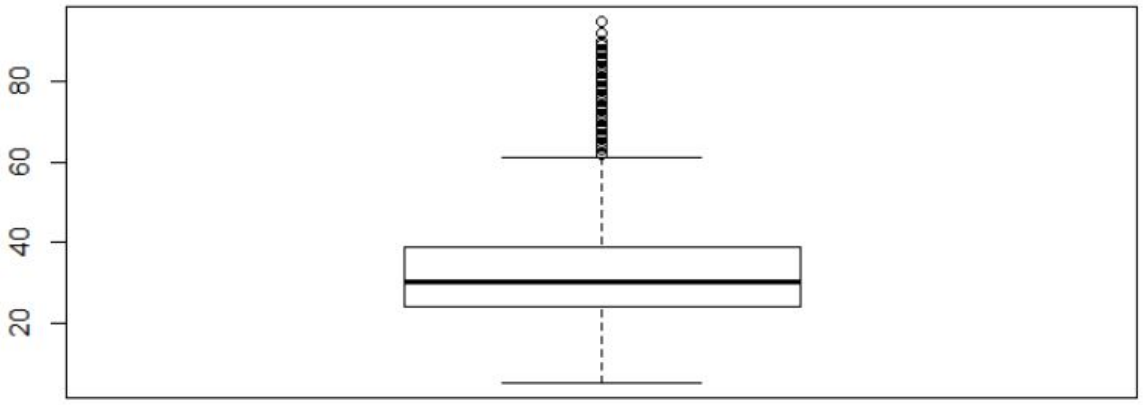
## Análisis de dataset de vehículos

```
    dia_ocu          mes_ocu        dia_sem_ocu     hora_ocu
 Min.   : 1.00   Min.   : 1.000   1: 7024     Min.   : 0.00
 1st Qu.: 8.00   1st Qu.: 4.000   2: 5963     1st Qu.: 8.00
 Median :16.00   Median : 7.000   3: 6384     Median :15.00
 Mean   :15.65   Mean   : 6.615   4: 7113     Mean   :13.62
 3rd Qu.:23.00   3rd Qu.:10.000   5: 7804     3rd Qu.:19.00
 Max.   :31.00   Max.   :12.000   6:10230     Max.   :99.00
                                  7:10442
   depto_ocu      area_geo_ocu     zona_ocu      sexo_per      edad_per
 1      :19005   1   :18192   99     :38843   1:42772   Min.   :  5.0
 5      : 4413   2   :28124   1      : 3494   2: 3055   1st Qu.: 26.0
 9      : 2572   NA's: 8644   12     : 1508   9: 9133   Median : 35.0
 10     : 2162                7      : 1444             Mean   :257.1
 17     : 2140                18     : 1035             3rd Qu.: 62.0
 18     : 2104                6      : 1024             Max.   :999.0
 (Other):22564                (Other): 7612
 estado_per     tipo_veh        color_veh       modelo_veh     tipo_eve
 1:18527    4      :16287   99     :14623   9999   :25716   1 :36558
 2: 5764    1      :12143   1      : 7816   9      : 5682   2 : 4865
 9:30669    3      : 8990   2      : 7754   6      : 3663   3 : 2356
            99     : 3477   5      : 7044   4      : 1832   4 : 1454
            2      : 3311   4      : 5371   3      : 1368   5 : 9684
            5      : 3219   3      : 4522   5      : 1039   13:   12
            (Other): 7533   (Other): 7830   (Other):15660   99:   31
 mayor_menor     num_hecho        marca_veh        g_edad_pil
 1   :39311   Min.   :    1   99     : 7598   11     :  922
 2   : 1099   1st Qu.: 1963   999    : 6307   3      :  877
 9   : 9150   Median : 3926   44     : 4082   4      :  827
 NA's: 5400   Mean   : 4234   69     : 2933   5      :  707
              3rd Qu.: 6115   32     : 1619   6      :  533
              Max.   :11618   (Other):21773   (Other): 1382
                              NA's   :10648   NA's   :49712
>
```

En la imagen anterior se observa el resumen de las variables que contiene el dataset de vehículos.

## Variables numéricas relevantes:






Se puede observar como en la figura anterior la media de edades de las persona involucrada en los accidentes es de aproximadamente de 30 años.

## Correlación entre variables variables numéricas

*Cuadro 4: Análisis de correlación entre las variables numéricas del dataset de Vehículos implicados*

| | dia_ocu | mes_ocu | dia_sem_ocu | hora_ocu | edad_per |
|---|---|---|---|---|---|
| dia_ocu | 1 | 0.008380382 | -0.005375771 | 0.004132864 | 0.012582701 |
| mes_ocu | 0.008380382 | 1 | 0.001893215 | -0.007756524 | -0.004758796 |
| dia_sem_ocu | -0.005375771 | 0.001893215 | 1 | -0.052063696 | -0.000263578 |
| hora_ocu | 0.004132864 | -0.007756524 | -0.052063696 | 1 | 0.001378908 |
| edad_per | 0.012582701 | -0.004758796 | -0.000263578 | 0.001378908 | 1 |

*Cuadro 5: Análisis de correlación entre las variables numéricas del dataset de personas implicadas*

| | dia | ano | hora | edad |
|---|---|---|---|---|
| dia | 1 | 0.019641 | 0.007021 | 0.009666 |
| ano | 0.019641 | 1 | -0.0099 | 0.070504 |
| hora | 0.007021 | -0.0099 | 1 | -0.00016 |
| edad | 0.009666 | 0.070504 | -0.00016 | 1 |

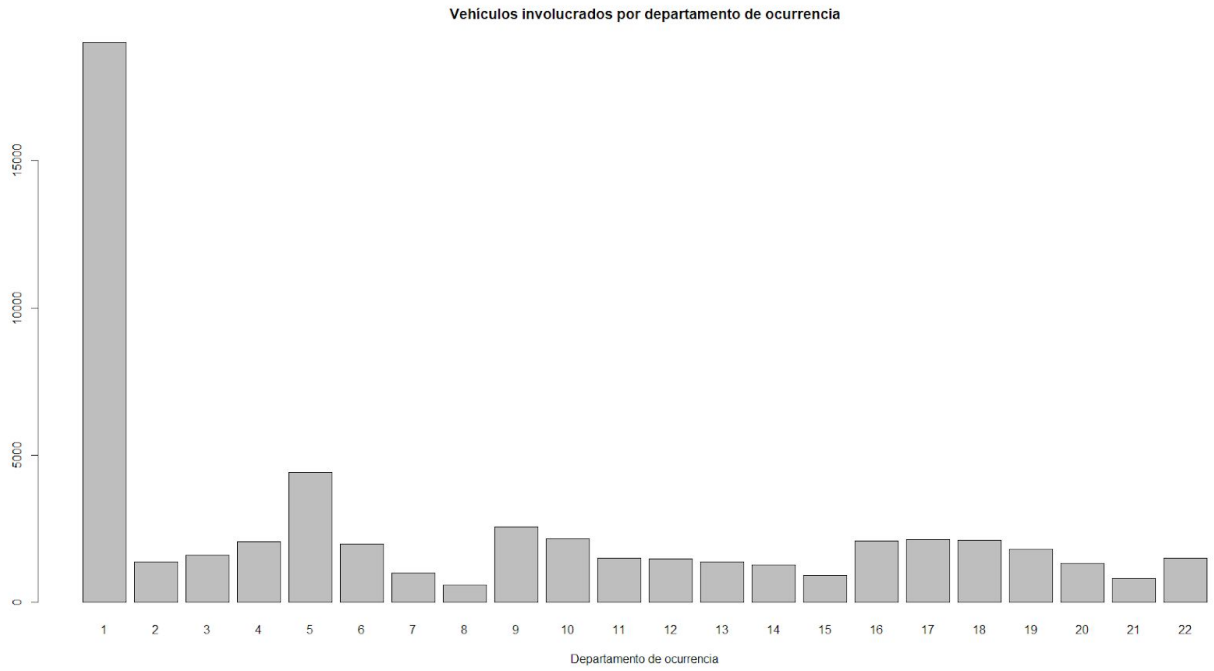
*Cuadro 6: Análisis de correlación entre las variables numéricas del dataset de hechos*

| | dia_ocu | ano_ocu | hora_ocu | edad_pil |
|---|---|---|---|---|
| dia_ocu | 1 | 0.008543277 | 0.005602882 | 0.01229476 |
| ano_ocu | 0.008543277 | 1 | 0.006037626 | 0.16411196 |
| hora_ocu | 0.005602882 | 0.006037626 | 1 | 0.0118738 |
| edad_pil | 0.012294764 | 0.164111957 | 0.011873805 | 1 |

Basados en las figuras anteriores, dado que ninguna de las variables tiene un número cercano a 1 se puede asegurar que no existe una correlación entre las variables, por lo tanto se pueden utilizar para realizar árboles de decisión en análisis posteriores.

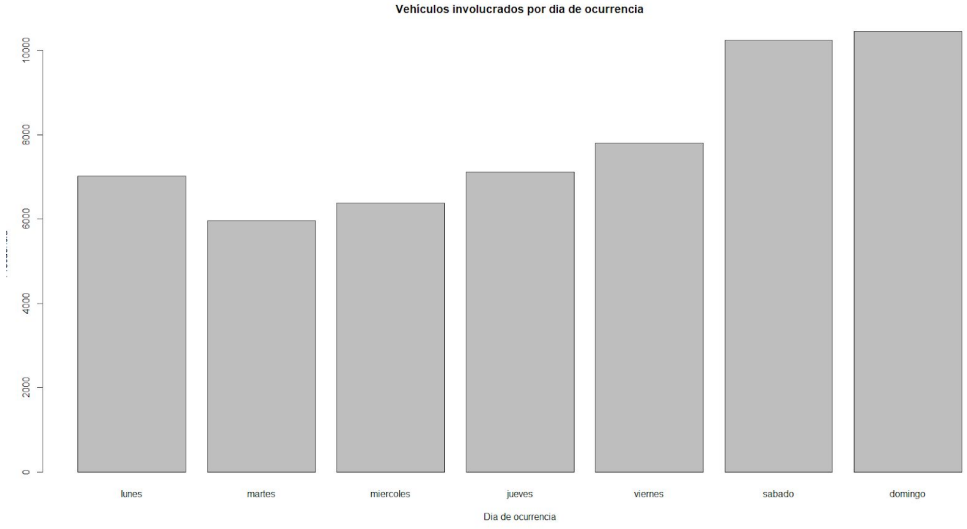
## Descripción de las variables categóricas Dataset Vehículos.

*Figura 1: vehículos involucrados por departamento*



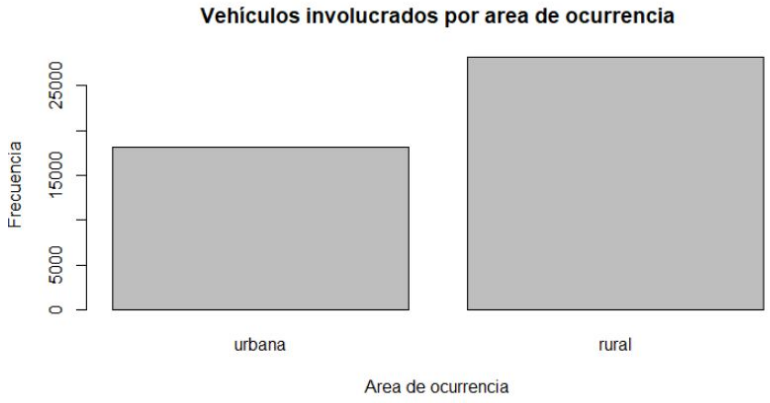

Se observa que la mayor cantidad de accidentes se da en el departamento de Guatemala seguido de Escuintla. Siendo el de menor cantidad de accidentes el de totonicapán.

*Figura 2: vehículos involucrados por dia de la semana*



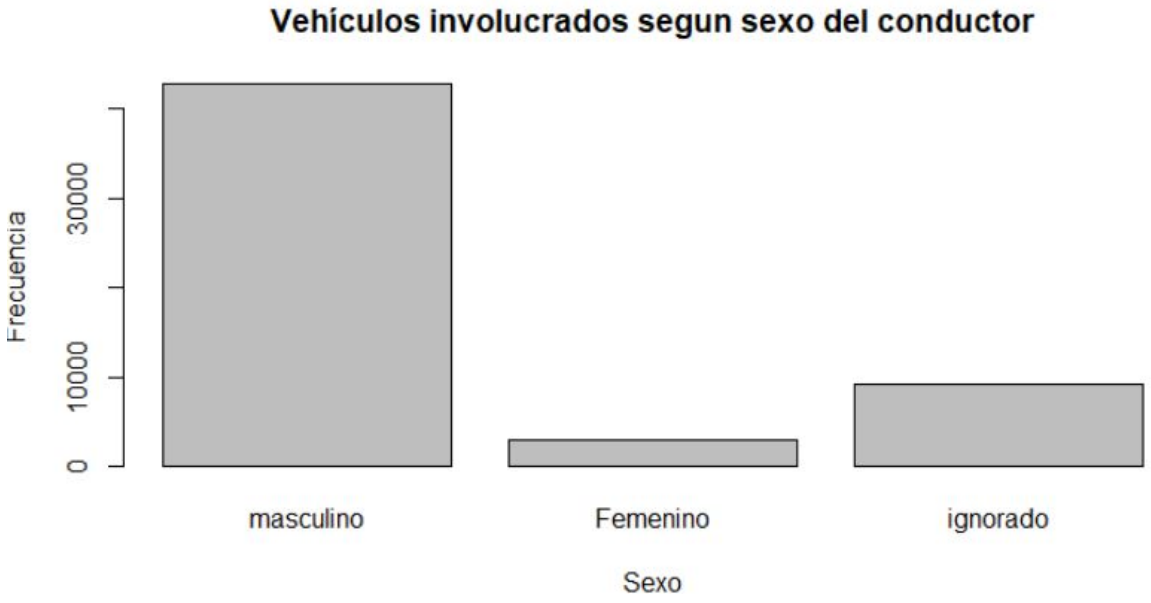

Los fines de semana son los días en que hay más accidentes.

*Figura 3: vehículos involucrados por área geográfica*



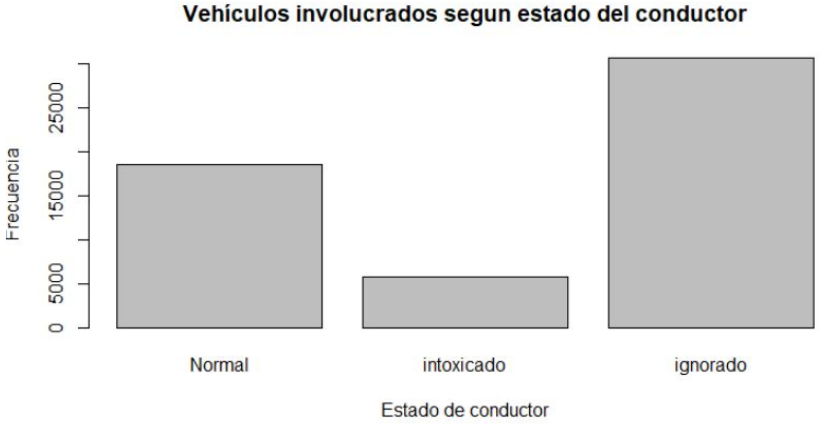

En el área rural sucedieron más accidentes que en los sitios rurales.

*Figura 4: vehículos involucrados según sexo del conductor*



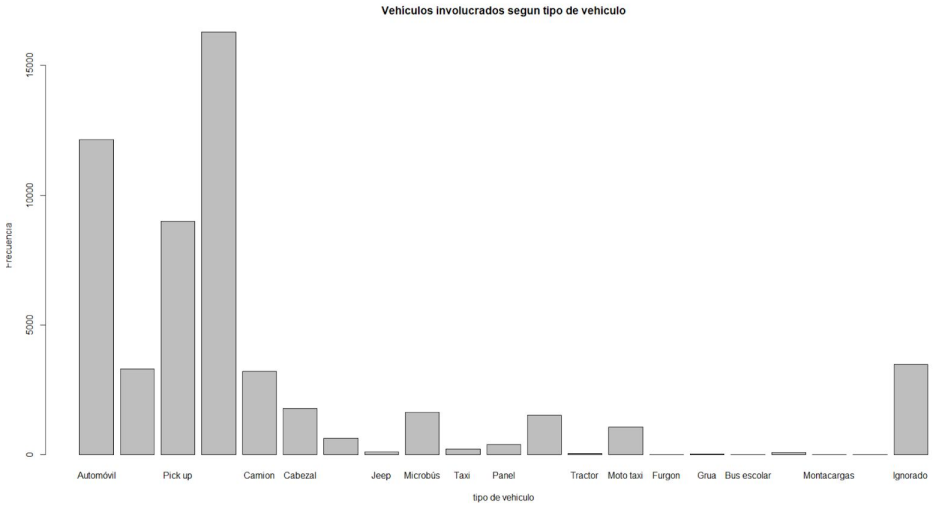

La mayor parte de percances le ocurren a personas de sexo masculino.

*Figura 5: vehículos involucrados según estado del conductor*



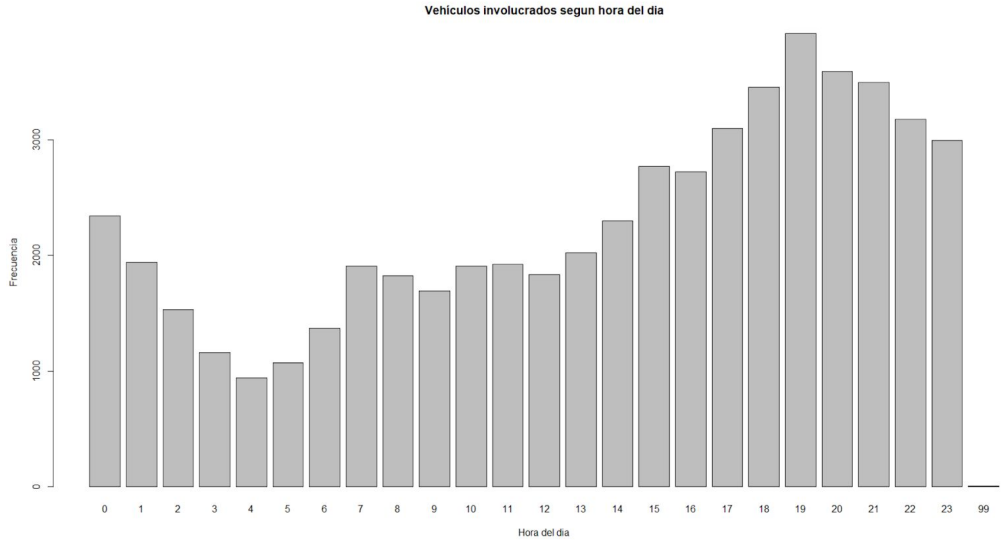
La mayor parte de los conductores no se tomó ninguno tipo de sustancia.

*Figura 6: vehículos involucrados según tipo de vehículo.*




Las motocicletas, fueron las que tuvieron más accidentes, seguidos de los vehículos.

*Figura 7: vehículos involucrados según hora del día*




Durante, las horas de la noche sucedieron la mayor parte de los accidentes, especialmente a las 19 horas, que es hora pico.

**Análisis de dataset de Hecho Tránsito**

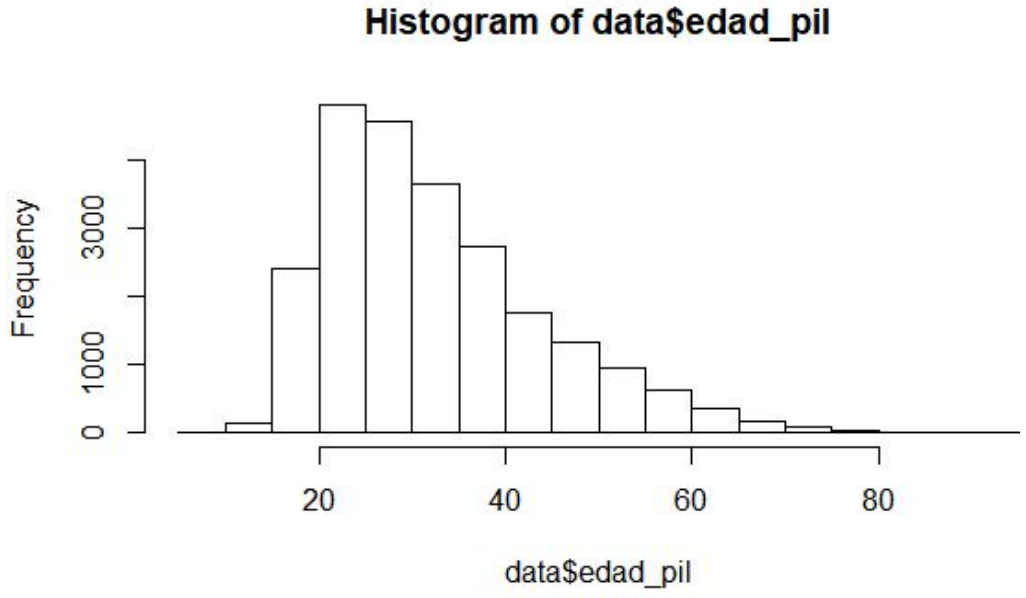
**Resumen**

```
   dia_ocu          mes_ocu          año_ocu      dia_sem_ocu
1       : 1828   12      : 4284   2016   :7964   1:5830
2       : 1621   8       : 3925   2015   :6854   2:5063
16      : 1611   4       : 3922   2013   :6324   3:5274
15      : 1558   3       : 3895   2017   :5879   4:5787
20      : 1550   7       : 3801   2014   :5651   5:6311
18      : 1538   10      : 3718   2009   :3528   6:8327
(Other):35524   (Other):21685   (Other):9030   7:8638
   hora_ocu        depto_ocu         tipo_veh        color_veh
Min.    : 0.00   1       :14890   4       :11105   99      :13265
1st Qu.: 8.00   5       : 3513   1       :10097   2       : 6605
Median :15.00   9       : 2096   3       : 8261   1       : 5985
Mean   :13.65   10      : 1847   99      : 3818   5       : 5107
3rd Qu.:19.00   16      : 1820   5       : 2686   4       : 4158
Max.   :99.00   17      : 1780   2       : 2667   3       : 3490
                (Other):19284   (Other): 6596   (Other): 6620
causa_acc     num_hecho          zona_ocu        modelo_veh
1 :22452   22      :   20   99      :29967   9999   :30681
2 : 5707   41      :   20   1       : 2337   2007   :  818
3 : 2998   62      :   20   7       : 1123   2008   :  741
4 : 1825   89      :   20   12      : 1094   2006   :  645
5 :12212   108     :   20   18      :  767   2012   :  605
13:   12   (Other):41943   (Other): 6414   2000   :  563
99:   24   NA's    : 3187   NA's    : 3528   (Other):11177
  marca_veh        mupio_ocu          g_hora       g_modelo_veh
99      : 7699   101     : 7394   Min.   :1.000   1:  101
999     : 4695   501     : 1224   1st Qu.:2.000   2: 1791
44      : 2953   901     :  829   Median :3.000   3: 4098
69      : 2127   1801    :  811   Mean   :2.861   4: 5539
21      : 1121   108     :  790   3rd Qu.:4.000   5: 3016
(Other):16770   (Other):24772   Max.   :4.000   6:30685
NA's    : 9865   NA's    : 9410   NA's   :11
```

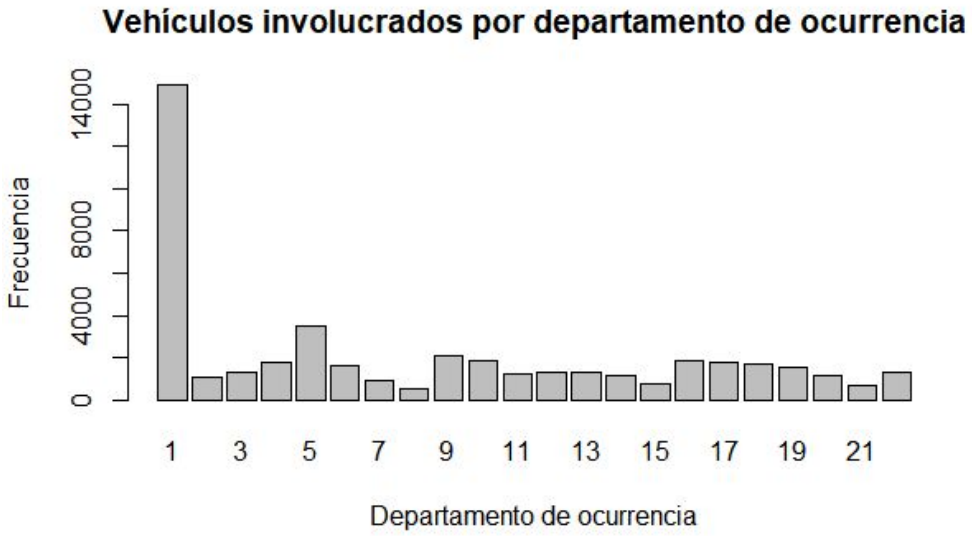
```
   g_hora_5      areag_ocu     sexo_pil        edad_pil
Min.   :1.000   1   :15375   1   :24587   Min.   :  5.0
1st Qu.:1.000   2   :23976   2   : 1500   1st Qu.: 26.0
Median :2.000   NA's: 5879   9   : 5300   Median : 36.0
Mean   :2.022                NA's:13843   Mean   :271.1
3rd Qu.:3.000                             3rd Qu.: 72.0
Max.   :3.000                             Max.   :999.0
NA's   :24533                             NA's   :13843
mayor_menor  estado_pil     corre_base    edad_quinquenales
1   :24520   1   :10549   1       :    2   18      : 2296
2   :  595   2   : 3757   2       :    2   6       :  934
9   : 5917   9   :17081   3       :    2   5       :  924
99  :  355   NA's:13843   4       :    2   7       :  699
NA's:13843                5       :    2   8       :  518
                          (Other): 5641   (Other): 1483
                          NA's   :39579   NA's   :38376

    g_edad
16      : 7733
4       : 4700
3       : 4416
5       : 3846
6       : 2933
(Other): 7759
NA's    :13843
```

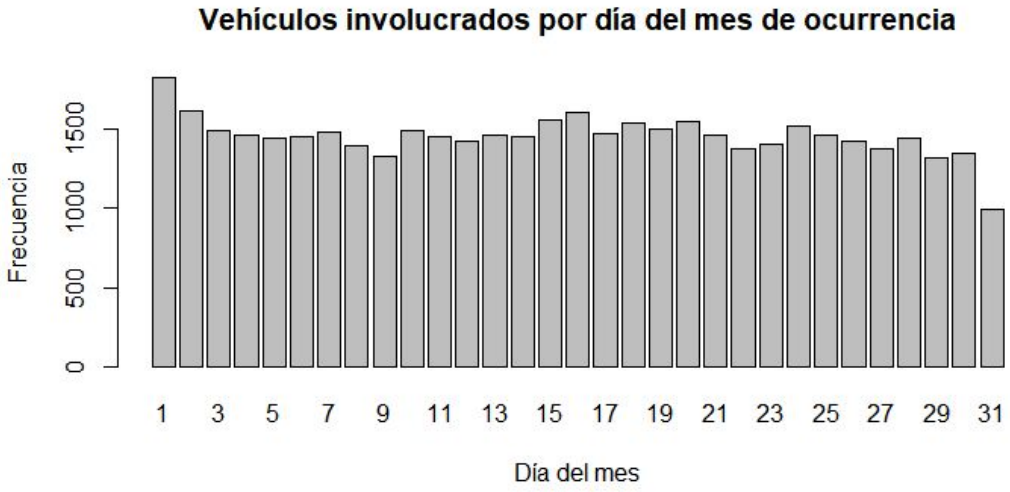
**Variables Numéricas**



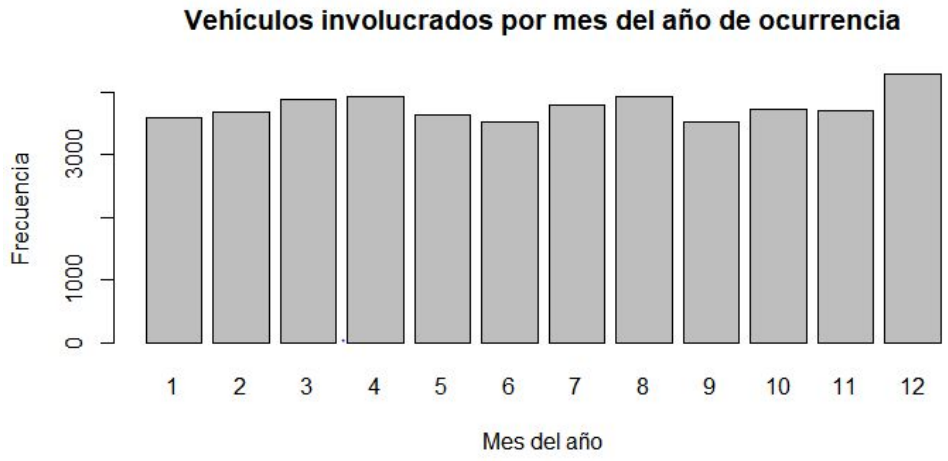
**Variables Categóricas**



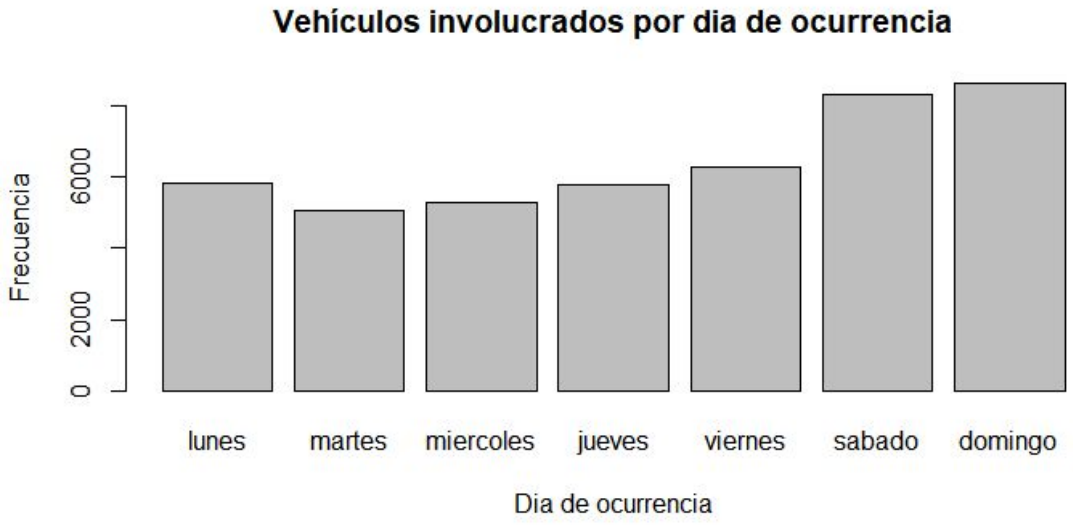

El departamento en el que ocurren más accidentes es el departamento # 1 que es el departamento de Guatemala.



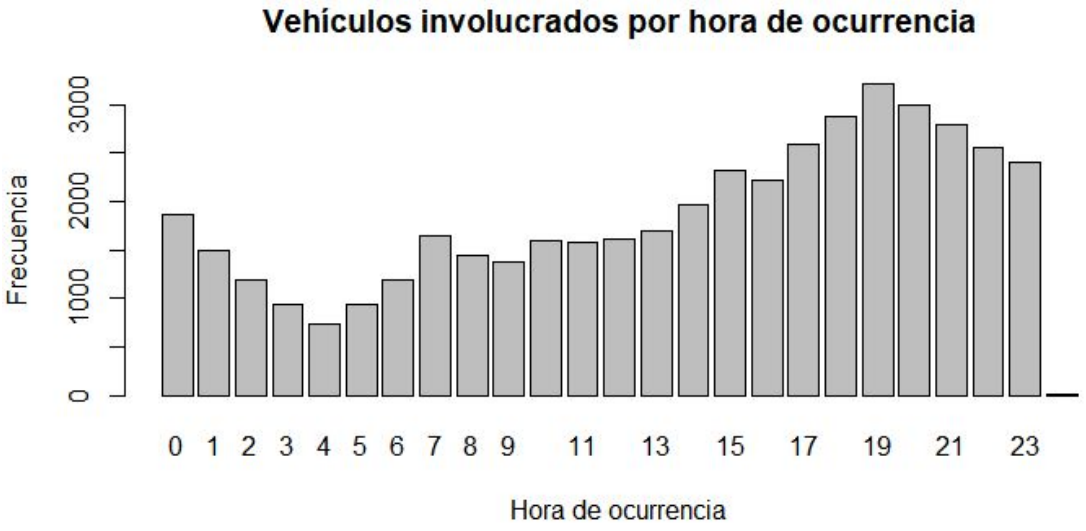

El día del mes en el que hay más accidentes es el primer día de cada mes.



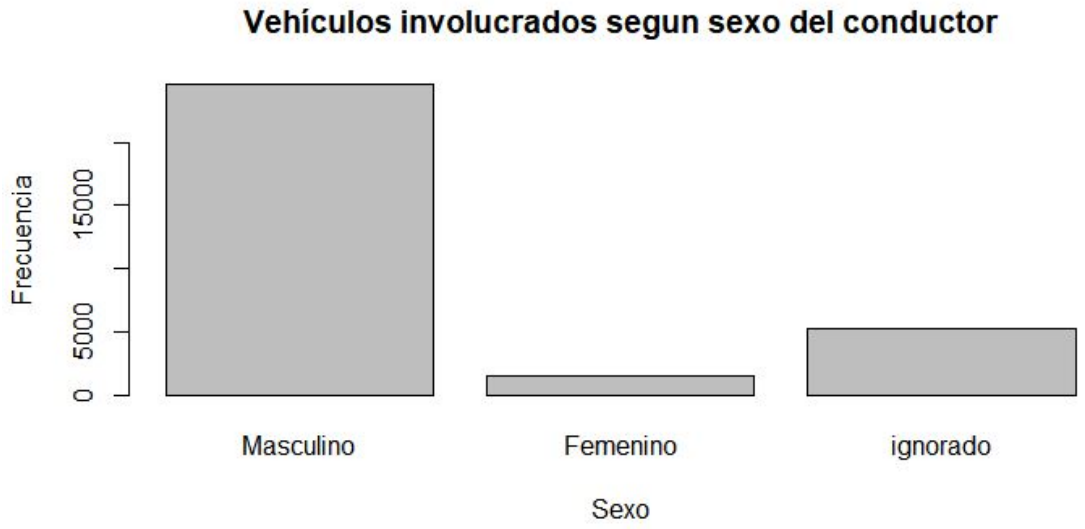

El mes del año en que más accidentes ocurren es el mes de diciembre.



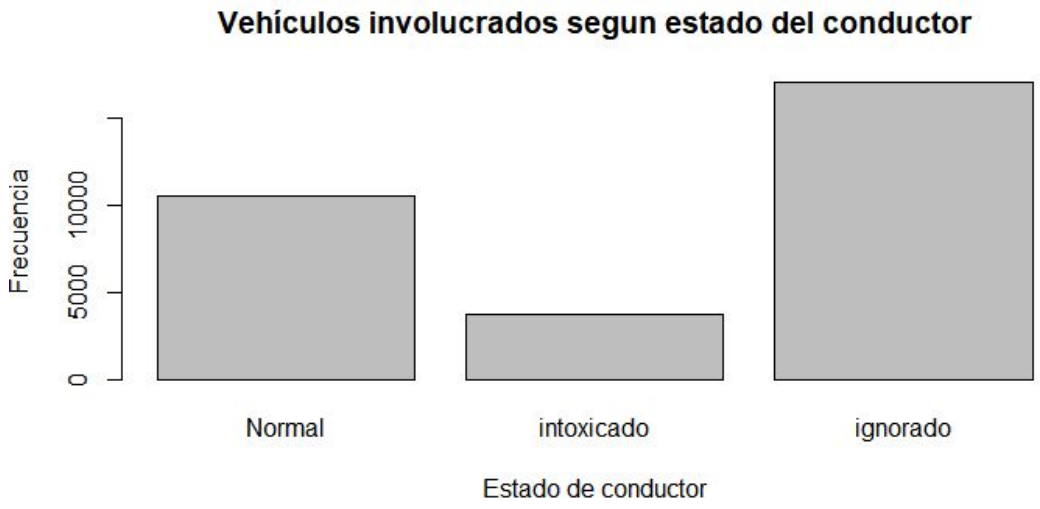

Los días de la semana en que más accidentes ocurren son sábado y domingo.




La hora en la que más accidentes ocurren está entre las 5 y 8 de la noche.




El sexo del conductor que más incurre es el masculino.




Los conductores que se encuentran en un estado normal son más propensos a sufrir un accidente.

## Análisis de dataset de Fallecidos Lesionados

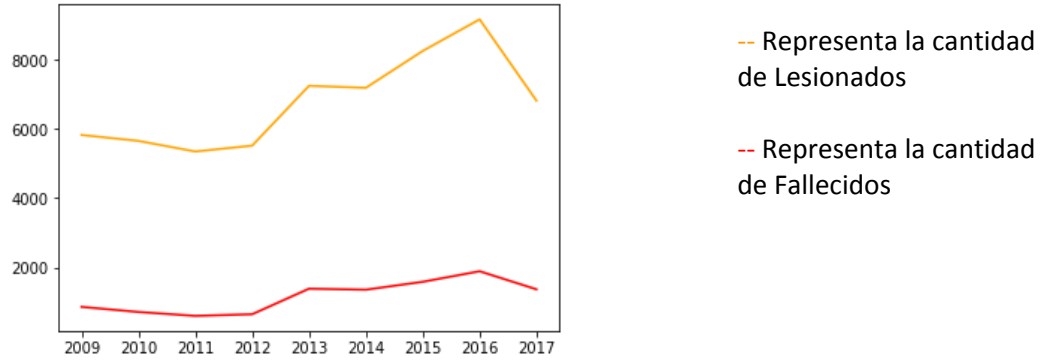
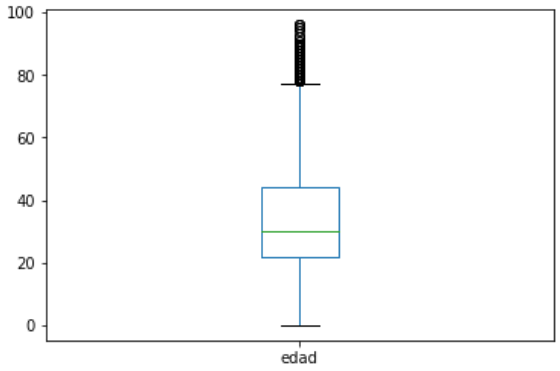
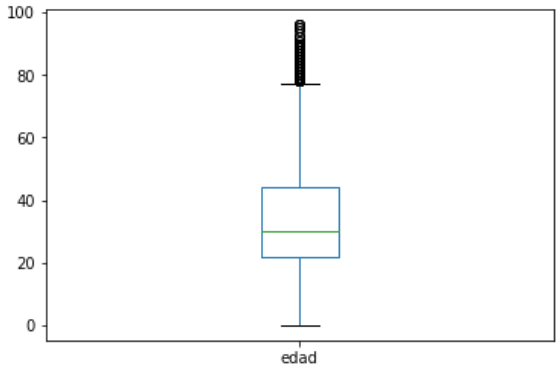



*Figura: Crecimiento/Disminución en los años de la cantidad de Lesionados y Fallecidos*

Se puede ver en la Figura No. que la cantidad de lesionados es mucho mayor que la cantidad de fallecidos en los accidentes de tránsito. A partir del 2016 la cantidad de ambos eventos a tenido una disminución abrupta, pero esta no se encuentra relacionada con el crecimiento vehicular. Los años en que más eventos de fallecidos y lesionados tuvieron fue en el 2016.




## Agrupamiento de variables

Los tres datasets que se tienen contienen en su mayoría información categórica, por lo tanto, no se puede utilizar un método de clustering que utilice la distancia Euclidiana para realizar los clusters porque solo acepta valores numéricos continuos. Se utilizará la distancia Gower para obtener la matriz de distancias.

*Distancia Gower*

La distancia Gower consiste en que, por cada tipo de variable se utilizará una métrica útil diferente que trabaje correctamente. Esta métrica se escalará para que tenga valores dentro de 0 y 1. Por último, se realiza una combinación lineal utilizando pesos para crear la matriz de distancia. Para las variables cuantitativas se utiliza la distancia Manhattan con los valores normalizados. Para valores ordinales, se les asigna un valor y luego se usa Manhattan. Por último, para las nominales, se usa el coeficiente de Dice que crea una columna binomial por cada subcategoría de cada variable nominal, luego calcula la proporción de apariciones por cada una.

*Algoritmo de clustering*

Ya con la matriz calculada, se procede a realizar el clustering. Para esto se utilizará PAM porque es más robusto para conjuntos de datos ruidosos y atípicos. Además, como resumen proporciona uno de los datos que representa al clustering, lo que nos ayuda a realizar un posterior análisis.

*Número de clusters*

Para obtener el número de clusters se utilizará el parámetro de silhouette para cada número de clusters probado en PAM.

## Clusters vehículos

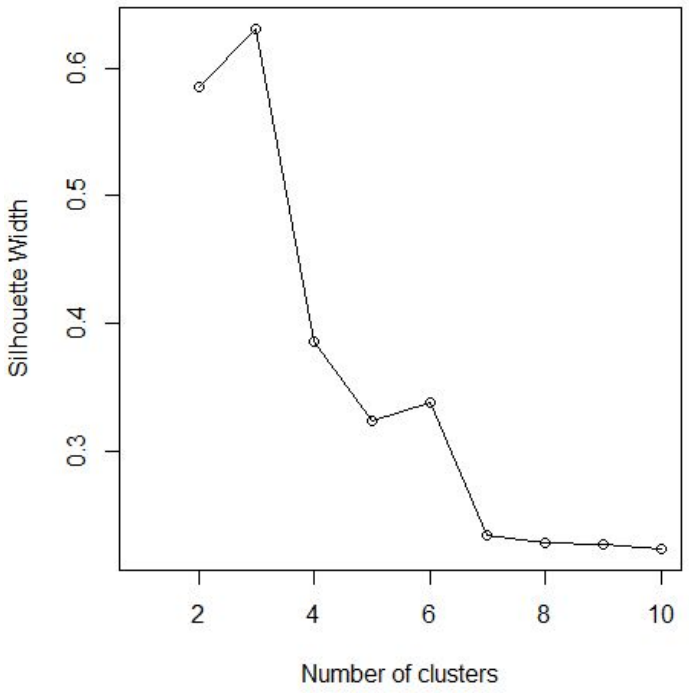



*Figura: Valor de Silhouette según la cantidad de clusters utilizando PAM*

Como se observa en la figura anterior, el número apropiado de clusters es de 3 porque se maximiza la silhouette. Por lo tanto se escoge 3 como número de clusters.

*Clusters para vehículos involucrados*

*[[1]]*

| dia_ocu | hora_ocu | edad_per | area_geo_ocu | sexo_per | cluster |
|---|---|---|---|---|---|
| 1 : 444 | 19 : 823 | Min. : 9.0 | 1 : 243 | 1:7376 | Min. :1 |
| 16 : 283 | 17 : 489 | 1st Qu.: 26.0 | 2 :6463 | 2: 411 | 1st Qu.:1 |
| 2 : 279 | 16 : 488 | Median : 35.0 | NA's:1090 | 9: 9 | Median :1 |
| 3 : 276 | 21 : 486 | Mean :114.3 | | | Mean :1 |
| 19 : 271 | 15 : 466 | 3rd Qu.: 46.0 | | | 3rd Qu.:1 |
| 26 : 271 | 20 : 460 | Max. :999.0 | | | Max. :1 |
| (Other):5972 | (Other):4584 | | | | |

*[[2]]*

| dia_ocu | hora_ocu | edad_per | area_geo_oc | u sexo_pe | r cluster |
|---|---|---|---|---|---|
| 15 : 396 | 18 : 651 | Min. : 10.00 | 1 :4555 | 1:5453 | Min. :2 |
| 18 : 235 | 23 : 383 | 1st Qu.: 23.00 | 2 : 294 | 2: 458 | 1st Qu.:2 |
| 17 : 221 | 20 : 361 | Median : 29.00 | NA's:1112 | 9: 50 | Median :2 |
| 13 : 216 | 21 : 360 | Mean : 99.23 | | | Mean :2 |
| 10 : 207 | 22 : 338 | 3rd Qu.: 38.00 | | | 3rd Qu.:2 |
| 3 : 202 | 17 : 322 | Max. :999.00 | | | Max. :2 |
| (Other):4484 | (Other):3546 | | | | |

*[[3]]*

| dia_ocu | hora_ocu | edad_per | area_geo_ocu | sexo_per | cluster |
|---|---|---|---|---|---|
| 20 : 160 | 20 : 321 | Min. : 16.0 | 1 : 666 | 1: 83 | Min. :3 |
| 11 : 109 | 19 : 213 | 1st Qu.:999.0 | 2 :1663 | 2: 65 | 1st Qu.:3 |
| 24 : 104 | 21 : 192 | Median :999.0 | NA's: 402 | 9:2583 | Median :3 |
| 4 : 101 | 22 : 167 | Mean :979.5 | | | Mean :3 |
| 27 : 100 | 23 : 151 | 3rd Qu.:999.0 | | | 3rd Qu.:3 |
| 28 : 99 | 18 : 129 | Max. :999.0 | | | Max. :3 |
| (Other):2058 | (Other):1558 | | | | |

*Figura: Descripción de cada cluster obtenido*

El primer cluster tiene vehículos que chocaron en su mayoría en el área rural y con mayoría de hombres, su media es de 26 años. El segundo cluster es de los vehículos que han chocado en el área urbana, teniendo una media de edad de 29 y mayoría de hombres. El último cluster contiene a los que no se les conoció su edad.

## Clusters personas implicadas

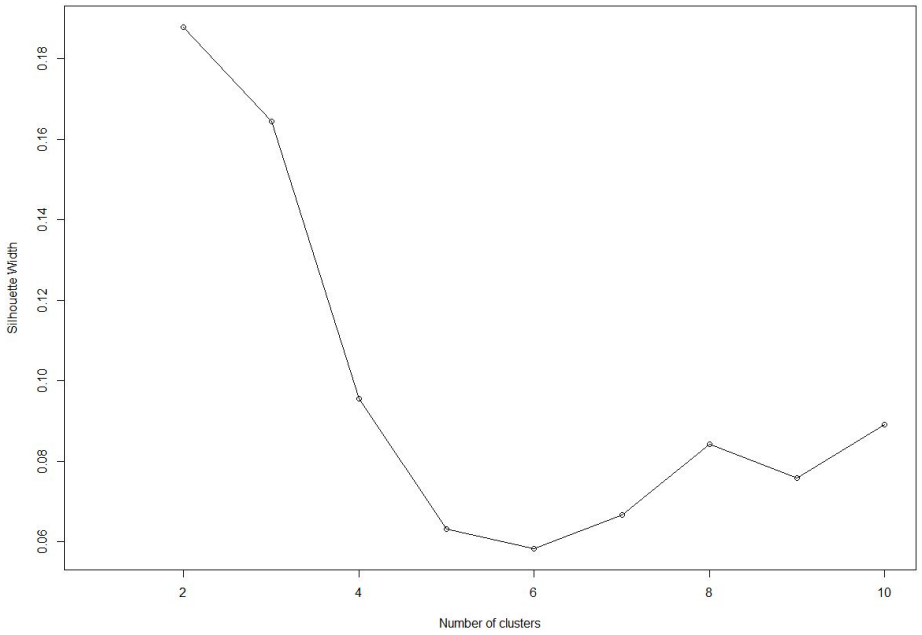



*Figura: Valor de Silhouette según la cantidad de clusters utilizando PAM*

Como se observa en la gráfica de silueta el valor que tendría la mejor cantidad de clusters sería 2. Sin embargo, esta sigue teniendo un valor muy pequeño, 0.16. Esto nos indica que los valores están demasiado dispersos y se debería probar con diferentes variables del dataset.

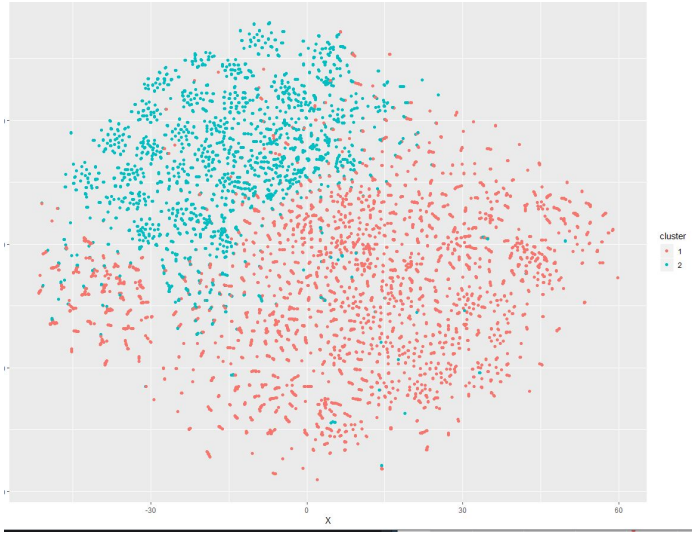



*Figura: Representación gráfica de los 2 clusters creados*

Como se observa en la Figura anterior, los valores se encuentran distribuidos de manera aparentemente distribuida.

```
[[1]]
    dia_ocu        hora_ocu       edad_pil     areag_ocu   sexo_pil
 1      : 307   19     : 587   Min.   :  5.0   1   : 168   1   :2914
 11     : 212   20     : 405   1st Qu.: 27.0   2   :4655   2   : 155
 3      : 209   17     : 370   Median : 39.0   NA's:1062   9   : 747
 26     : 206   21     : 369   Mean   :311.1               NA's:2069
 17     : 203   22     : 329   3rd Qu.:999.0
 20     : 201   16     : 307   Max.   :999.0
 (Other):4547   (Other):3518   NA's   :2069
    cluster
 Min.   :1
 1st Qu.:1
 Median :1
 Mean   :1
 3rd Qu.:1
 Max.   :1
```

```
[[2]]
    dia_ocu        hora_ocu       edad_pil     areag_ocu   sexo_pil
 10     : 215   18     : 376   Min.   : 12.0   1   :2856   1   :2001
 18     : 120   23     : 195   1st Qu.: 25.0   2   : 181   2   : 158
 19     : 120   22     : 189   Median : 32.0   NA's: 124   9   : 278
 13     : 117   21     : 188   Mean   :197.5               NA's: 724
 21     : 116   17     : 178   3rd Qu.: 48.0
 24     : 112   20     : 168   Max.   :999.0
 (Other):2361   (Other):1867   NA's   :724
    cluster
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2
```

*Figura: Descripción de cada cluster obtenido*

En esta figura se muestra que no existe una variable que diferencie a los clusters entre sí.

**Clusters de hechos sucedidos**

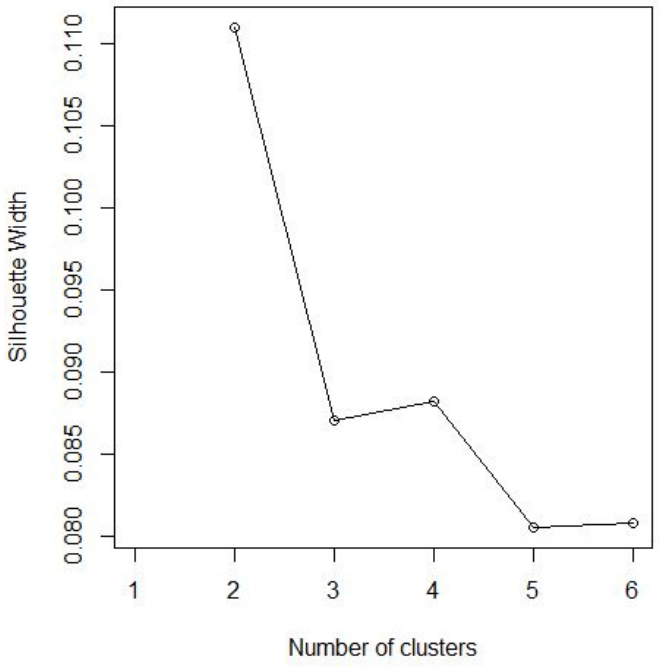


*Figura: Gráfica de silhouette para las personas implicadas en el accidente*

Se escoge k = 2 para los clusters. Pero, todos presentan un coeficiente de silhouette muy bajo. Por lo que los resultados obtenidos no son significativos. Con las variables utilizadas no se pueden crear clusters. Por lo tanto, se termina aquí este análisis.

# Conclusiones y Hallazgos

1. Se encontró, que los fines de semana y las noches son las más propensas para que se sufra un accidente de tránsito.
2. Los vehículos con más accidentes reportados son los de motocicletas, debido a su gran popularidad.
3. Se pueden analizar los accidentes de forma separada dependiendo si fueron en un área rural o urbana.
4. Existen más hombres involucrados en los accidentes que mujeres
5. El día con más accidentes es el 1ro de cada mes
6. Las personas involucradas en los accidentes se encuentran entre 20 y 30 años

## Siguiente paso a seguir

Considerando los hallazgos de la información, se podría plantear qué tipo de modelo se utilizará para realizar predicciones. Por la gran cantidad de variables categóricas, se podría utilizar Random Forest o Naive Bayes (de los que se han aprendido hasta ahora).