

Memoria Competición Kaggel Preprocesamiento

Francisco Pérez Hernández

09/03/2017

Contents

1	Introducción al problema y a Kaggel	2
1.1	Lectura del dataset accidentes	2
1.2	Primera prueba con un modelo	4
1.3	Creación del archivo de salida y subida a kaggel	5
2	Análisis del dataset	6
2.1	Eliminación de valores perdidos	6
2.2	Prueba del modelo con eliminación de variables	10
3	Vusualización del dataset	10
3.1	Análisis de las variables actuales	10
3.2	Análisis de variables eliminadas sin valores perdidos	20
4	Visión preeliminar de los datos	26
5	Imputación de valores perdidos	27
5.1	Imputación de variables	28
5.2	Prueba del modelo con imputación de valores perdidos	29
6	Detección de anomalías	30
6.1	Uso del paquete outliers	30
6.2	Paquete mvoutlier	35
6.3	Eliminación de valores anómalos	35
6.4	Prueba del modelo con imputación de valores perdidos	38
7	Transformación de los datos	39
7.1	Transformando los datos	40
7.2	Prueba del modelo con transformación de los datos	40
8	Discretización	40
9	Selección de características	40
9.1	Paquete FSelector	42

1 Introducción al problema y a Kaggel

Lo primero que se pretende realizar en este apartado es leer el dataset que nos han dado y realizar una subida a la plataforma Kaggel para obtener una primera puntuación. Mi usuario en Kaggel es “PacoPollos”.

1.1 Lectura del dataset accidentes

Vamos a leer tanto los archivos de train como test dados.

```
accidentes.train.original <- read.csv("accidentes-kaggle.csv")
accidentes.test.original <- read.csv("accidentes-kaggle-test.csv")
```

Una vez leídos vamos a realizar un summary para ver como están compuestos los datos.

```
summary(accidentes.train.original)
```

```
##          ANIO              MES              HORA              DIASEMANA
## Min.      :2008      Julio      : 2757      14      : 1965      DOMINGO   :3597
## 1st Qu.:2009      Junio       : 2649      19      : 1847      JUEVES    :4351
## Median :2010      Mayo        : 2605      13      : 1823      LUNES     :4349
## Mean     :2010      Octubre    : 2600      17      : 1749      MARTES    :4343
## 3rd Qu.:2012      Septiembre: 2491      18      : 1726      MIERCOLES:4394
## Max.     :2013      Diciembre  : 2448      12      : 1713      SABADO    :4000
##          (Other)      :14452      (Other):19179      VIERNES   :4968
##          PROVINCIA              COMUNIDAD_AUTONOMA              ISLA
## Barcelona: 6238      Cataluna              :8208      NO_ES_ISLA :28476
## Madrid    : 4735      Madrid, Comunidad de:4735      MALLORCA   : 608
## Valencia  : 1658      Andalucia              :4412      TENERIFE   : 436
## Sevilla   : 977      Comunitat Valenciana:2653      GRAN CANARIA: 199
## Cadiz     : 887      Pais Vasco              :1594      IBIZA       : 117
## Girona    : 814      Castilla y Leon          :1505      LANZAROTE   : 53
## (Other)   :14693      (Other)                  :6895      (Other)     : 113
## TOT_VICTIMAS      TOT_MUERTOS      TOT_HERIDOS_GRAVES      TOT_HERIDOS_LEVES
## Min.      : 1.000      Min.      :0.00000      Min.      :0.0000      Min.      : 0.00
## 1st Qu.: 1.000      1st Qu.:0.00000      1st Qu.:0.0000      1st Qu.: 1.00
## Median : 1.000      Median :0.00000      Median :0.0000      Median : 1.00
## Mean     : 1.429      Mean     :0.02447      Mean     :0.1453      Mean     : 1.26
## 3rd Qu.: 2.000      3rd Qu.:0.00000      3rd Qu.:0.0000      3rd Qu.: 1.00
## Max.     :19.000      Max.     :7.00000      Max.     :9.0000      Max.     :18.00
##
## TOT_VEHICULOS_IMPLICADOS      ZONA              ZONA_AGRUPADA
## Min.      : 1.000      CARRETERA :13278      VIAS INTERURBANAS:13335
## 1st Qu.: 1.000      TRAVESIA  : 241      VIAS URBANAS      :16667
## Median : 2.000      VARIANTE  : 57
## Mean     : 1.738      ZONA URBANA:16426
## 3rd Qu.: 2.000
## Max.     :21.000
##
##          CARRETERA
## A-7      : 294
## A-2      : 278
## AP-7     : 229
## N-340    : 229
## A-4      : 184
```

```

## (Other):12098
## NA's :16690
##
## RED_CARRETERA
## OTRAS TITULARIDADES : 318
## TITULARIDAD AUTONOMICA : 3890
## TITULARIDAD ESTATAL : 4021
## TITULARIDAD MUNICIPAL :19077
## TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
##
## TIPO_VIA
## OTRO TIPO :15527
## VIA CONVENCIONAL:10044
## AUTOVIA : 2941
## AUTOPISTA : 723
## CAMINO VECINAL : 519
## RAMAL DE ENLACE : 101
## (Other) : 147
##
## TRAZADO_NO_INTERSEC
## CURVA FUERTE CON MARCA Y SIN VELOCIDAD MARCADA: 559
## CURVA FUERTE CON MARCA Y VELOCIDAD MARCADA : 872
## CURVA FUERTE SIN MARCAR : 481
## CURVA SUAVE : 2875
## ES_INTERSECCION :11038
## RECTA :14177
##
## TIPO_INTERSEC
## EN T O Y : 3350
## EN X O + : 4714
## ENLACE DE ENTRADA : 421
## ENLACE DE SALIDA : 223
## GIRATORIA : 2006
## NO_ES_INTERSECCION:18983
## OTROS : 305
##
## ACOND_CALZADA
## CARRIL CENTRAL DE ESPERA : 193
## NADA ESPECIAL : 4645
## OTRO TIPO : 791
## PASO PARA PEATONES O ISLETAS EN CENTRO DE VIA PRINCIPAL: 397
## RAQUETA DE GIRO IZQUIERDA : 109
## SOLO ISLETAS O PASO PARA PEATONES : 168
## NA's :23699
##
## PRIORIDAD SUPERFICIE_CALZADA
## NINGUNA (SOLO NORMA) :13495 SECA Y LIMPIA :25236
## SEMAFORO : 1778 MOJADA : 3895
## SEAL DE STOP : 1750 OTRO TIPO : 327
## SOLO MARCAS VIALES : 1659 UMBRIA : 165
## SEAL DE CEDA EL PASO: 1629 GRAVILLA SUELTA: 150
## (Other) : 1569 ACEITE : 83
## NA's : 8122 (Other) : 146
##
## LUMINOSIDAD FACTORES_ATMOSFERICOS
## CREPUSCULO : 1330 BUEN TIEMPO :25852
## NOCHE: ILUMINACION INSUFICIENTE: 1067 LLOVIZNANDO : 2524
## NOCHE: ILUMINACION SUFICIENTE : 4793 OTRO : 715

```

```
## NOCHE: SIN ILUMINACION      : 1815    LLUVIA FUERTE: 499
## PLENO DIA                   :20997    VIENTO FUERTE: 156
##                             NIEBLA LIGERA: 83
##                             (Other)    : 173
## VISIBILIDAD_RESTRINGIDA      OTRA_CIRCUNSTANCIA
## SIN RESTRICCION              :16982    NINGUNA      :24967
## CONFIGURACION DEL TERRENO: 989      OTRA         : 942
## OTRA_CAUSA                   : 491     OBRAS        : 263
## FACTORES ATMOSFERICOS       : 374     FUERTE DESCENSO : 227
## EDIFICIOS                   : 229     CAMBIO DE RASANTE: 100
## (Other)                     : 252     (Other)      : 264
## NA's                        :10685    NA's         : 3239
## ACERAS                      DENSIDAD_CIRCULACION MEDIDAS_ESPECIALES
## NO HAY ACERA:21416 CONGESTIONADA: 308 CARRIL REVERSIBLE : 17
## SI HAY ACERA: 5437 DENSA : 1479 HABILITACION ARCEN: 8
## NA's : 3149 FLUIDA :17505 NINGUNA MEDIDA :21024
## NA's :10710 OTRA MEDIDA : 278
## NA's : 8675
##
## TIPO_ACCIDENTE
## Atropello : 3642
## Colision_Obstaculo: 952
## Colision_Vehiculos:16520
## Otro : 1807
## Salida_Via : 6013
## Vuelco : 1068
##
```

Vemos como las variables TTO_VICTIMAS, TOT_MUERTOS, TOT_HERIDOS_GRAVES, TOT_HERIDOS_LEVES y TOT_VEHICULOS_IMPLICADOS son las únicas variables numéricas, por lo que nos quedaremos con ellas para la primera prueba, junto con la variable clasificadora TIPO_ACCIDENTE.

```
accidentes.train.solo.numericos <- accidentes.train.original[,c(8,9,10,11,12,30)]
accidentes.test.solo.numericos <- accidentes.test.original[,c(8,9,10,11,12)]
```

1.2 Primera prueba con un modelo

Lo primero es, con las variables numéricas únicamente, voy a realizar un primer modelo, que será un árbol, para predecir la clase del conjunto de test y comprobar el funcionamiento de Kaggle al no tener experiencia anterior.

```
set.seed(1234)
ct1 <- ctree(TIPO_ACCIDENTE ~., accidentes.train.solo.numericos)
testPred1 <- predict(ct1, newdata = accidentes.test.solo.numericos)
```

Por lo que ya tenemos el conjunto de test predicho. Además el árbol creado tendría la siguiente estructura:

```
ct1

##
## Conditional inference tree with 14 terminal nodes
##
## Response: TIPO_ACCIDENTE
## Inputs: TOT_VICTIMAS, TOT_MUERTOS, TOT_HERIDOS_GRAVES, TOT_HERIDOS_LEVES, TOT_VEHICULOS_IMPLICADOS
```

```

## Number of observations: 30002
##
## 1) TOT_VEHICULOS_IMPLICADOS <= 1; criterion = 1, statistic = 14488.658
## 2) TOT_VICTIMAS <= 1; criterion = 1, statistic = 329.362
## 3) TOT_HERIDOS_GRAVES <= 0; criterion = 1, statistic = 38.228
## 4) TOT_HERIDOS_LEVES <= 0; criterion = 0.996, statistic = 21.181
## 5)* weights = 256
## 4) TOT_HERIDOS_LEVES > 0
## 6)* weights = 7696
## 3) TOT_HERIDOS_GRAVES > 0
## 7)* weights = 1476
## 2) TOT_VICTIMAS > 1
## 8) TOT_VICTIMAS <= 2; criterion = 1, statistic = 47.735
## 9)* weights = 1605
## 8) TOT_VICTIMAS > 2
## 10)* weights = 550
## 1) TOT_VEHICULOS_IMPLICADOS > 1
## 11) TOT_HERIDOS_LEVES <= 1; criterion = 1, statistic = 99.886
## 12) TOT_HERIDOS_LEVES <= 0; criterion = 1, statistic = 49.242
## 13)* weights = 1276
## 12) TOT_HERIDOS_LEVES > 0
## 14) TOT_VICTIMAS <= 1; criterion = 1, statistic = 34.382
## 15) TOT_VEHICULOS_IMPLICADOS <= 3; criterion = 1, statistic = 28.319
## 16) TOT_VEHICULOS_IMPLICADOS <= 2; criterion = 0.999, statistic = 24.207
## 17)* weights = 10133
## 16) TOT_VEHICULOS_IMPLICADOS > 2
## 18)* weights = 924
## 15) TOT_VEHICULOS_IMPLICADOS > 3
## 19)* weights = 254
## 14) TOT_VICTIMAS > 1
## 20) TOT_VEHICULOS_IMPLICADOS <= 3; criterion = 0.965, statistic = 15.891
## 21)* weights = 370
## 20) TOT_VEHICULOS_IMPLICADOS > 3
## 22)* weights = 21
## 11) TOT_HERIDOS_LEVES > 1
## 23) TOT_VEHICULOS_IMPLICADOS <= 4; criterion = 0.994, statistic = 20.095
## 24) TOT_VEHICULOS_IMPLICADOS <= 2; criterion = 0.998, statistic = 22.592
## 25)* weights = 4183
## 24) TOT_VEHICULOS_IMPLICADOS > 2
## 26)* weights = 1124
## 23) TOT_VEHICULOS_IMPLICADOS > 4
## 27)* weights = 134

```

1.3 Creación del archivo de salida y subida a kaggle

Vamos a escribir la salida del primer modelo para ver su puntuación en Kaggle.

```

salida.primer.modelo <- as.matrix(testPred1)
salida.primer.modelo <- cbind(c(1:(dim(salida.primer.modelo)[1])), salida.primer.modelo)
colnames(salida.primer.modelo) <- c("Id", "Prediction")
write.table(salida.primer.modelo, file="predicciones/PrimeraPrediccion.txt", sep="," , quote = F, row.names = F)

```

Por lo que ya tenemos un fichero con la salida del conjunto de test. Lo único que tendremos que modificar es la primera línea del archivo para añadir “Id, Prediction”. El resultado de este primer modelo para la

competición de Kaggle, subido el 11/02/2017 a las 19:54, con un total de 5 personas entregadas, se ha quedado en la posición 3 con una puntuación del 0.73246.

#	Δ3d	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Luis Suárez	0.82948	2	Fri, 10 Feb 2017 19:54:58
2	↓1	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
3	new	PacoPollos	0.73246	1	Sat, 11 Feb 2017 18:51:32
4	↓1	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
5	↓1	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10

Figure 1: Primera puntuación obtenida en Kaggle

2 Análisis del dataset

Una vez realizada la primera prueba en Kaggle, vamos a analizar con detalle el dataset que nos han dado.

2.1 Eliminación de valores perdidos

Anteriormente en el summary, hemos visto que hay variables con valores perdidos, ya que por ejemplo, en la variable CARRETERA uno de los valores que más se repite es NA's. Por lo tanto, vamos a analizar que variables contienen valores perdidos.

```
porcentaje.de.valores.perdidos.por.columna.train <- apply(accidentes.train.original,2,function(x) sum(is.na(x)))
columnas.train.con.valores.perdidos <- (porcentaje.de.valores.perdidos.por.columna.train > 0)
columnas.train.con.valores.perdidos
```

```
##          ANIO          MES          HORA
##          FALSE          FALSE          FALSE
##          DIASEMANA          PROVINCIA          COMUNIDAD_AUTONOMA
##          FALSE          FALSE          FALSE
##          ISLA          TOT_VICTIMAS          TOT_MUERTOS
##          FALSE          FALSE          FALSE
##          TOT_HERIDOS_GRAVES          TOT_HERIDOS_LEVES          TOT_VEHICULOS_IMPLICADOS
##          FALSE          FALSE          FALSE
##          ZONA          ZONA_AGRUPADA          CARRETERA
##          FALSE          FALSE          TRUE
##          RED_CARRETERA          TIPO_VIA          TRAZADO_NO_INTERSEC
##          FALSE          FALSE          FALSE
##          TIPO_INTERSEC          ACOND_CALZADA          PRIORIDAD
##          FALSE          TRUE          TRUE
##          SUPERFICIE_CALZADA          LUMINOSIDAD          FACTORES_ATMOSFERICOS
##          FALSE          FALSE          FALSE
##          VISIBILIDAD_RESTRINGIDA          OTRA_CIRCUNSTANCIA          ACERAS
##          TRUE          TRUE          TRUE
##          DENSIDAD_CIRCULACION          MEDIDAS_ESPECIALES          TIPO_ACCIDENTE
##          TRUE          TRUE          FALSE
```

Por lo que tenemos que las variables con valores perdidos son: CARRETERA, ACOND_CALZADA, PRIORIDAD, VISIBILIDAD_RESTRINGIDA, OTRA_CIRCUNSTANCIA, ACERAS, DENSIDAD_CIRCULACION y MEDIDAS_ESPECIALES. Veamos el resumen para esas variables.

```
summary(accidentes.train.original[c("CARRETERA","ACOND_CALZADA","PRIORIDAD", "VISIBILIDAD_RESTRINGIDA",
```

```
##      CARRETERA
## A-7      : 294
## A-2      : 278
## AP-7     : 229
## N-340    : 229
## A-4      : 184
## (Other):12098
## NA's     :16690
##
##                                ACOND_CALZADA
## CARRIL CENTRAL DE ESPERA      : 193
## NADA ESPECIAL                 : 4645
## OTRO TIPO                     : 791
## PASO PARA PEATONES O ISLETAS EN CENTRO DE VIA PRINCIPAL: 397
## RAQUETA DE GIRO IZQUIERDA    : 109
## SOLO ISLETAS O PASO PARA PEATONES : 168
## NA's                         :23699
##
##          PRIORIDAD          VISIBILIDAD_RESTRINGIDA
## NINGUNA (SOLO NORMA) :13495 SIN RESTRICCION      :16982
## SEMAFORO             : 1778 CONFIGURACION DEL TERRENO: 989
## SEÑAL DE STOP        : 1750 OTRA_CAUSA          : 491
## SOLO MARCAS VIALES   : 1659 FACTORES ATMOSFERICOS : 374
## SEÑAL DE CEDA EL PASO: 1629 EDIFICIOS           : 229
## (Other)              : 1569 (Other)            : 252
## NA's                 : 8122 NA's              :10685
##
##          OTRA_CIRCUNSTANCIA          ACERAS          DENSIDAD_CIRCULACION
## NINGUNA      :24967 NO HAY ACERA:21416 CONGESTIONADA: 308
## OTRA         : 942 SI HAY ACERA: 5437 DENSA      : 1479
## OBRAS        : 263 NA's           : 3149 FLUIDA     :17505
## FUERTE DESCENSO : 227 NA's           :10710
## CAMBIO DE RASANTE: 100
## (Other)       : 264
## NA's         : 3239
##
##          MEDIDAS_ESPECIALES
## CARRIL REVERSIBLE : 17
## HABILITACION ARCEN: 8
## NINGUNA MEDIDA    :21024
## OTRA MEDIDA       : 278
## NA's              : 8675
##
##
```

Donde podemos ver que el valor más pequeño de NA's es para la variable ACERAS con 3149 instancias con valores perdidos, lo que sería un 10,49% de los datos. Un 25% de los datos de este train serían unas 7500 instancias, por lo que las variables que tienen más del 25% de valores perdidos son: CARRETERA, ACOND_CALZADA, PRIORIDAD, VISIBILIDAD_RESTRINGIDA, DENSIDAD_CIRCULACION y MEDIDAS_ESPECIALES. O lo que es lo mismo, me quedo con las variables OTRA_CIRCUNSTANCIA y ACERAS, del anterior grupo. Pero además voy a comenzar eliminando esas variables ya que a mi juicio pueden no tener demasiada importancia.

```

primeras.variables.eliminadas <- c("CARRETERA", "ACOND_CALZADA", "PRIORIDAD", "VISIBILIDAD_RESTRINGIDA",
accidentes.train.sin.variables.1 <- accidentes.train.original[,-c(15,20,21,25,26,27,28,29)]
accidentes.train.variables.eliminadas <- accidentes.train.original[,c(15,20,21,25,26,27,28,29)]

```

Por lo que guardo en una variable las variables que he eliminado, y creo mi dataset sin variables con valores NA. Hago lo mismo para el test:

```

accidentes.test.sin.variables.1 <- accidentes.test.original[,-c(15,20,21,25,26,27,28,29)]
accidentes.test.variables.eliminadas <- accidentes.test.original[,c(15,20,21,25,26,27,28,29)]
accidentes.test.variables.eliminadas.copia <- accidentes.test.variables.eliminadas

```

Pensemos ahora que variables restantes pueden ser no interesantes.

```
summary(accidentes.train.sin.variables.1)
```

```

##          ANIO          MES          HORA          DIASEMANA
## Min.   :2008   Julio    : 2757   14    : 1965   DOMINGO   :3597
## 1st Qu.:2009   Junio    : 2649   19    : 1847   JUEVES    :4351
## Median :2010   Mayo     : 2605   13    : 1823   LUNES     :4349
## Mean   :2010   Octubre  : 2600   17    : 1749   MARTES    :4343
## 3rd Qu.:2012   Septiembre: 2491   18    : 1726   MIERCOLES:4394
## Max.   :2013   Diciembre : 2448   12    : 1713   SABADO    :4000
##          (Other) :14452   (Other):19179   VIERNES   :4968
##          PROVINCIA          COMUNIDAD_AUTONOMA          ISLA
## Barcelona: 6238   Cataluna      :8208   NO_ES_ISLA :28476
## Madrid    : 4735   Madrid, Comunidad de:4735   MALLORCA   : 608
## Valencia  : 1658   Andalucia      :4412   TENERIFE   : 436
## Sevilla   : 977   Comunitat Valenciana:2653   GRAN CANARIA: 199
## Cadiz     : 887   Pais Vasco     :1594   IBIZA      : 117
## Girona    : 814   Castilla y Leon :1505   LANZAROTE  : 53
## (Other)   :14693   (Other)        :6895   (Other)    : 113
## TOT_VICTIMAS TOT_MUERTOS TOT_HERIDOS_GRAVES TOT_HERIDOS_LEVES
## Min.   : 1.000   Min.   :0.00000   Min.   :0.0000   Min.   : 0.00
## 1st Qu.: 1.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 1.00
## Median : 1.000   Median :0.00000   Median :0.0000   Median : 1.00
## Mean   : 1.429   Mean   :0.02447   Mean   :0.1453   Mean   : 1.26
## 3rd Qu.: 2.000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 1.00
## Max.   :19.000   Max.   :7.00000   Max.   :9.0000   Max.   :18.00
##
## TOT_VEHICULOS_IMPLICADOS          ZONA          ZONA_AGRUPADA
## Min.   : 1.000          CARRETERA :13278   VIAS INTERURBANAS:13335
## 1st Qu.: 1.000          TRAVESIA  : 241   VIAS URBANAS     :16667
## Median : 2.000          VARIANTE  : 57
## Mean   : 1.738          ZONA URBANA:16426
## 3rd Qu.: 2.000
## Max.   :21.000
##
##
##                                RED_CARRETERA
## OTRAS TITULARIDADES          : 318
## TITULARIDAD AUTONOMICA      : 3890
## TITULARIDAD ESTATAL         : 4021
## TITULARIDAD MUNICIPAL       :19077
## TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
##

```



```

##          TIPO_VIA
## OTRO TIPO      :15527
## VIA CONVENCIONAL:10044
## AUTOVIA       : 2941
## AUTOPISTA     :  723
## CAMINO VECINAL :  519
## RAMAL DE ENLACE :  101
## (Other)       :  147
##
##          TRAZADO_NO_INTERSEC
## CURVA FUERTE CON MARCA Y SIN VELOCIDAD MARCADA:  559
## CURVA FUERTE CON MARCA Y VELOCIDAD MARCADA   :  872
## CURVA FUERTE SIN MARCAR                      :  481
## CURVA SUAVE                                  : 2875
## ES_INTERSECCION                             :11038
## RECTA                                        :14177
##
##          TIPO_INTERSEC          SUPERFICIE_CALZADA
## EN T O Y      : 3350  SECA Y LIMPIA :25236
## EN X O +      : 4714  MOJADA       : 3895
## ENLACE DE ENTRADA :  421  OTRO TIPO   :  327
## ENLACE DE SALIDA  :  223  UMBRIA     :  165
## GIRATORIA       : 2006  GRAVILLA SUELTA:  150
## NO_ES_INTERSECCION:18983  ACEITE      :   83
## OTROS           :  305  (Other)     :  146
##
##          LUMINOSIDAD          FACTORES_ATMOSFERICOS
## CREPUSCULO      : 1330  BUEN TIEMPO :25852
## NOCHE: ILUMINACION INSUFICIENTE: 1067  LLOVIZNANDO : 2524
## NOCHE: ILUMINACION SUFICIENTE : 4793  OTRO       :  715
## NOCHE: SIN ILUMINACION        : 1815  LLUVIA FUERTE:  499
## PLENO DIA                :20997  VIENTO FUERTE:  156
##
##          NIEBLA LIGERA:  83
##          (Other)     :  173
##
##          TIPO_ACCIDENTE
## Atropello      : 3642
## Colision_Obstaculo:  952
## Colision_Vehiculos:16520
## Otro           : 1807
## Salida_Via     : 6013
## Vuelco         : 1068
##

```

Podemos pensar que otras de las variables que puede que no nos sean de mucha utilidad pueden ser: ANIO, MES, HORA, DIASEMANA, PROVINCIA, COMUNIDAD_AUTONOMA, ISLA, ZONA_AGRUPADA, TIPO_VIA, TRAZADO_NO_INTERSEC, TIPO_INTERSEC, SUPERFICIE_CALZADA y LUMINOSIDAD. Ya que muchas de estas variables podrían no ser de vital importancia, de primera mano, para la obtención de la predicción del tipo de accidente. Por lo tanto, vamos a eliminarlas de momento para agilizar los modelos primeros.

```

segundas.variables.eliminadas <- c("ANIO", "MES", "HORA", "DIASEMANA", "PROVINCIA", "COMUNIDAD_AUTONOMA",
accidentes.train.sin.variables.2 <- accidentes.train.sin.variables.1[,-c(1,2,3,4,5,6,7,14,16,17,18,19,20)]
accidentes.train.variables.eliminadas <- cbind(accidentes.train.variables.eliminadas ,accidentes.train.variables.2)
accidentes.test.sin.variables.2 <- accidentes.test.sin.variables.1[,-c(1,2,3,4,5,6,7,14,16,17,18,19,20)]
accidentes.test.variables.eliminadas <- cbind(accidentes.test.sin.variables.2 ,accidentes.test.variables.1)

```

2.2 Prueba del modelo con eliminación de variables

Hagamos por lo tanto una prueba de como afecta la inclusión de estas variables con respecto a la primera prueba realizada.

```
set.seed(1234)
ct2 <- ctree(TIPO_ACCIDENTE ~., accidentes.train.sin.variables.2)
testPred2 <- predict(ct2, newdata = accidentes.test.sin.variables.2)
```

Por lo que ya tenemos el conjunto de test predicho. Además el árbol creado tendría la siguiente estructura:

```
#ct2
```

Vamos a escribir la salida del modelo para ver su puntuación en Kaggel.

```
salida.segundo.modelo <- as.matrix(testPred2)
salida.segundo.modelo <- cbind(c(1:(dim(salida.segundo.modelo)[1])), salida.segundo.modelo)
colnames(salida.segundo.modelo) <- c("Id", "Prediction")
write.table(salida.segundo.modelo, file="predicciones/SegundaPrediccion.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 17/02/2017 a las 17:51, con un total de 14 personas entregadas, se ha quedado en la posición 9 con una puntuación del 0.81891.

3 Vusualización del dataset

Como no se ha hecho antes, y debería ser uno de los primeros pasos a realizar, vamos a realizar una visualización del dataset.

3.1 Análisis de las variables actuales

Vamos a ver el comportamiento de nuestras variables con respecto al TIPO_ACCIDENTE, a ver que relación pueden tener.

```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_VICTIMAS , y = TIPO_A
```


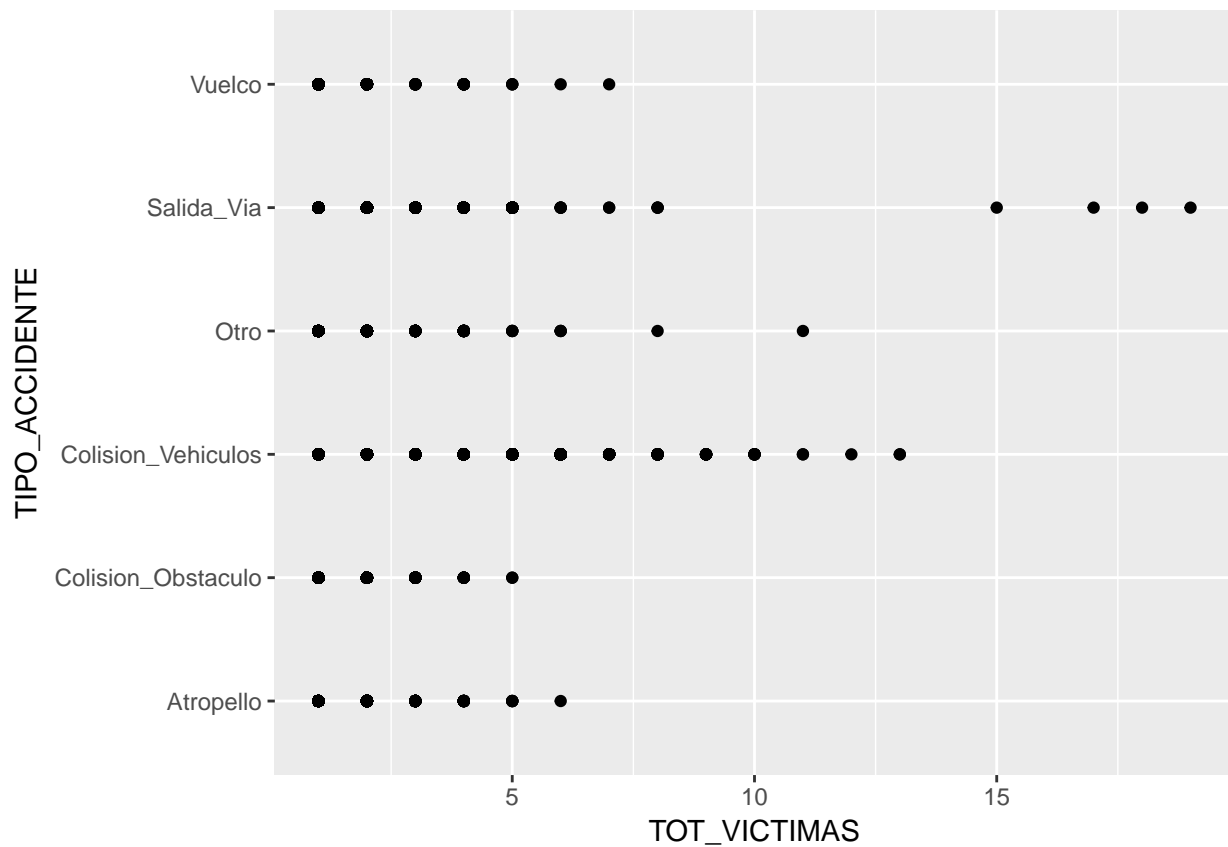
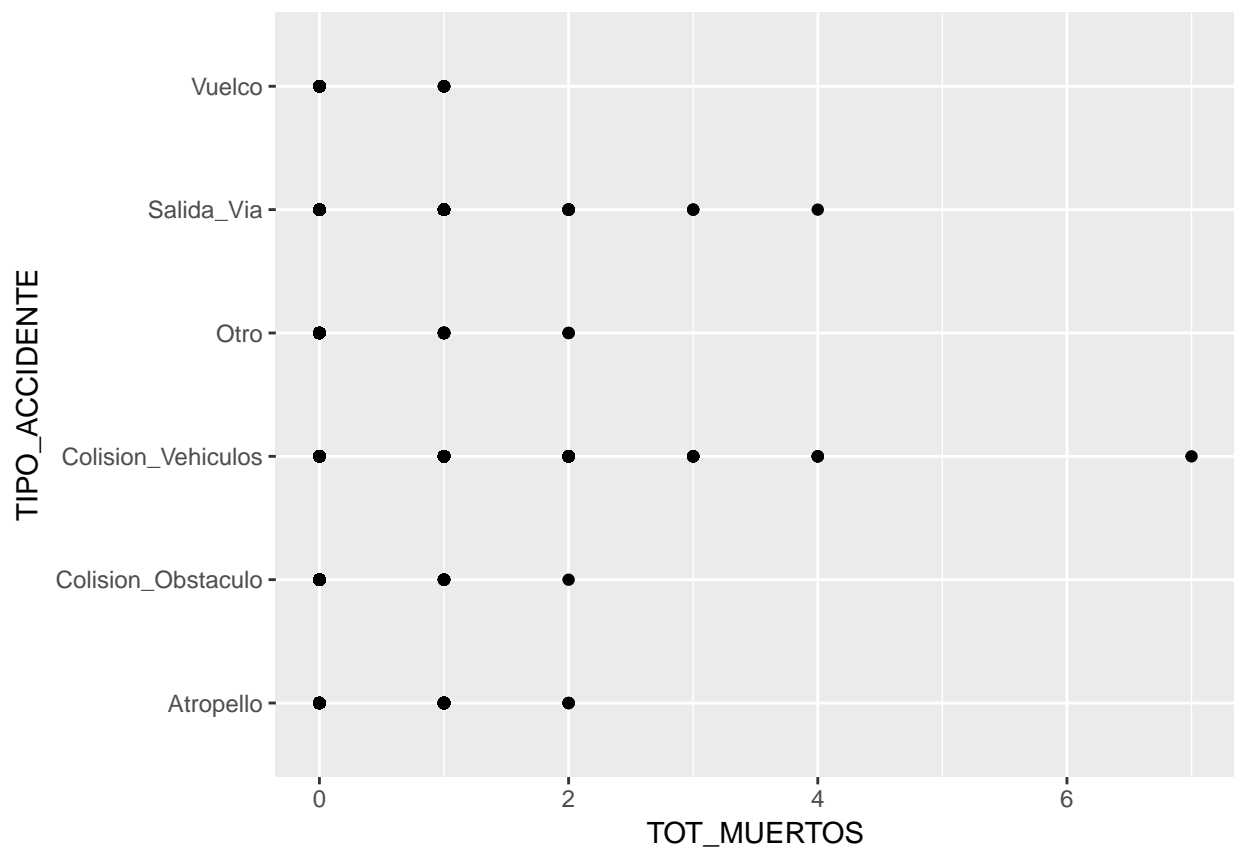
#	Δ5d	Team Name	Score 	Entries	Last Submission UTC (Best – Last Submission)
1	new	Anabel Gómez	0.83175	9	Fri, 17 Feb 2017 11:34:17 (-19.6h)
2	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
3	new	Jonathan Espinosa	0.82671	8	Thu, 16 Feb 2017 12:28:22
4	new	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
5	new	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
6	new	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
7	new	WhiteShadow	0.82247	3	Thu, 16 Feb 2017 13:06:30
8	↓5	PacoPollos	0.81891	2	Fri, 17 Feb 2017 16:50:29
<p>Your Best Entry ↑</p> <p>Top Ten!</p> <p>You made the top ten by improving your score by 0.08645.</p> <p>You just moved up 1 position on the leaderboard. Tweet this!</p>					
9	↓7	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
10	new	Jorge Jimena	0.73246	1	Fri, 17 Feb 2017 02:57:20
11	↓7	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
12	↓7	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
13	new	Xisco Fauli	0.48735	2	Wed, 15 Feb 2017 23:16:45
14	new	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 2: Segunda puntuación obtenida en Kaggel

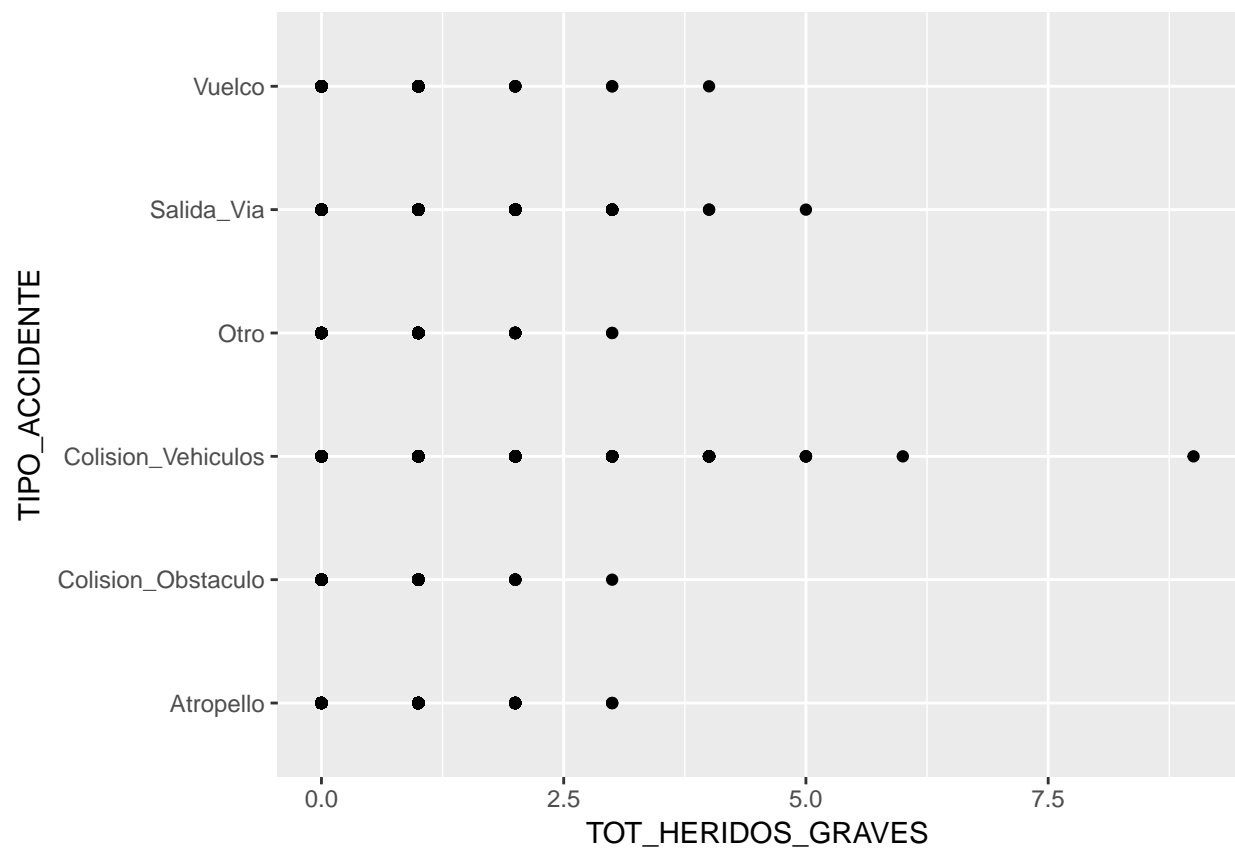


Podemos ver como para a partir de 10 victimas, el accidente suele ser o una colisión de vehículos, salida de vía, o muy pocas veces otro tipo de accidente. Por lo que puede ser una relación interesante.

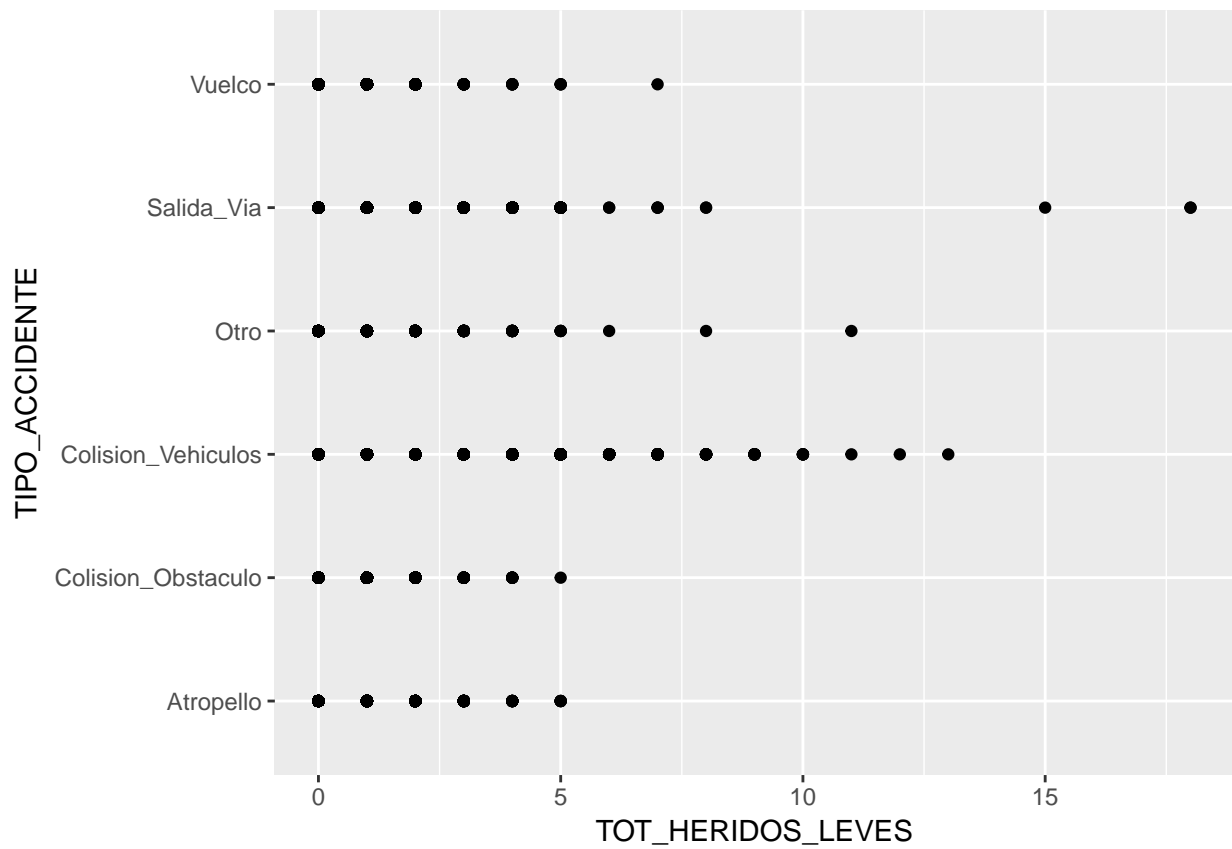
```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_MUERTOS , y = TIPO_A
```



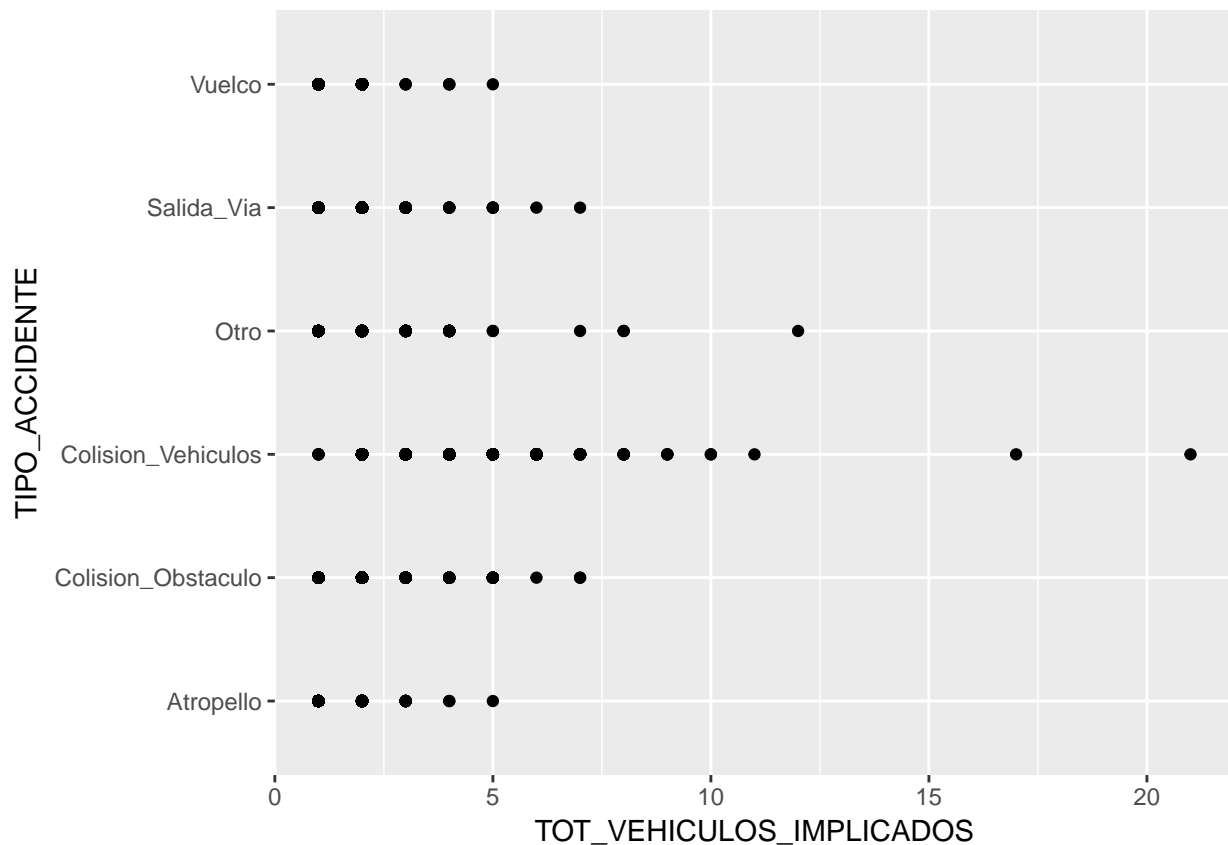
```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_HERIDOS_GRAVES , y =
```



```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_HERIDOS_GRAVES , y = TIPO_ACCIDENTE))
```

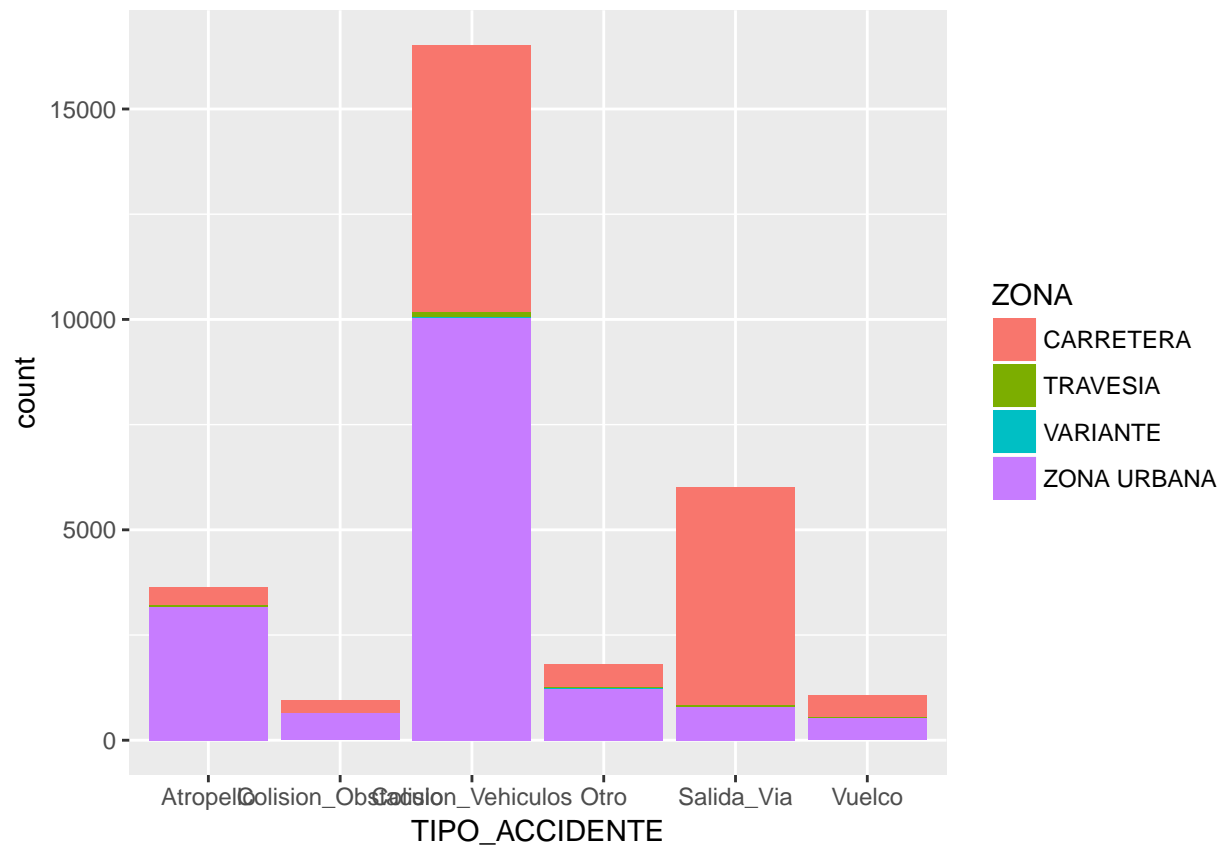


```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_VEHICULOS_IMPLICADOS
```



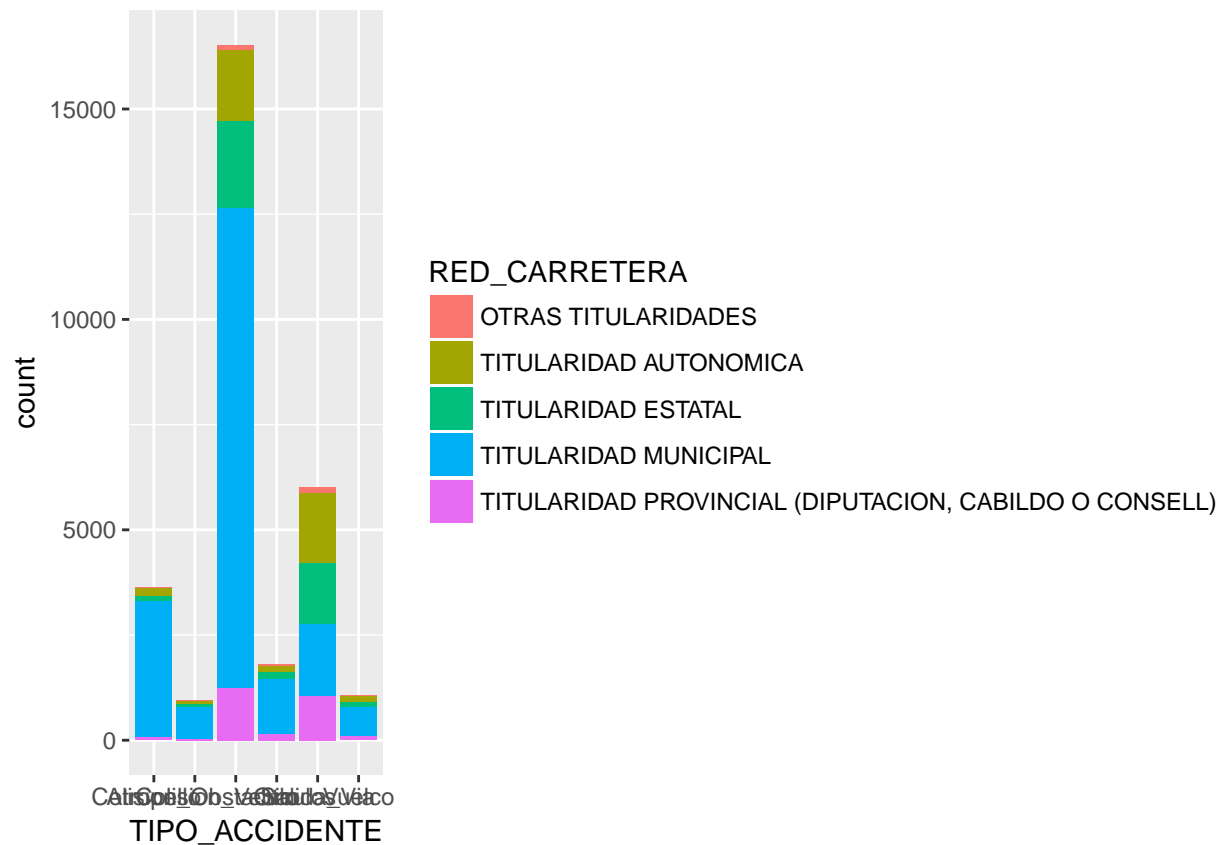
Normalmente a partir de 3 muertos, el accidente es una colisión de vehículos o una salida de vía. Si hay más de 3 heridos graves, suele ser colisión de vehículos, salida de vía o vuelco. A partir de 6 heridos leves el accidente es una colisión, una salida de vía, un vuelco o otro accidente. A partir de 6 vehículos implicados, los accidentes suelen ser colisiones, salida de vía u otro tipo. Por lo que ya tenemos varias relaciones que podrían ser representadas en un árbol.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=ZONA))
```

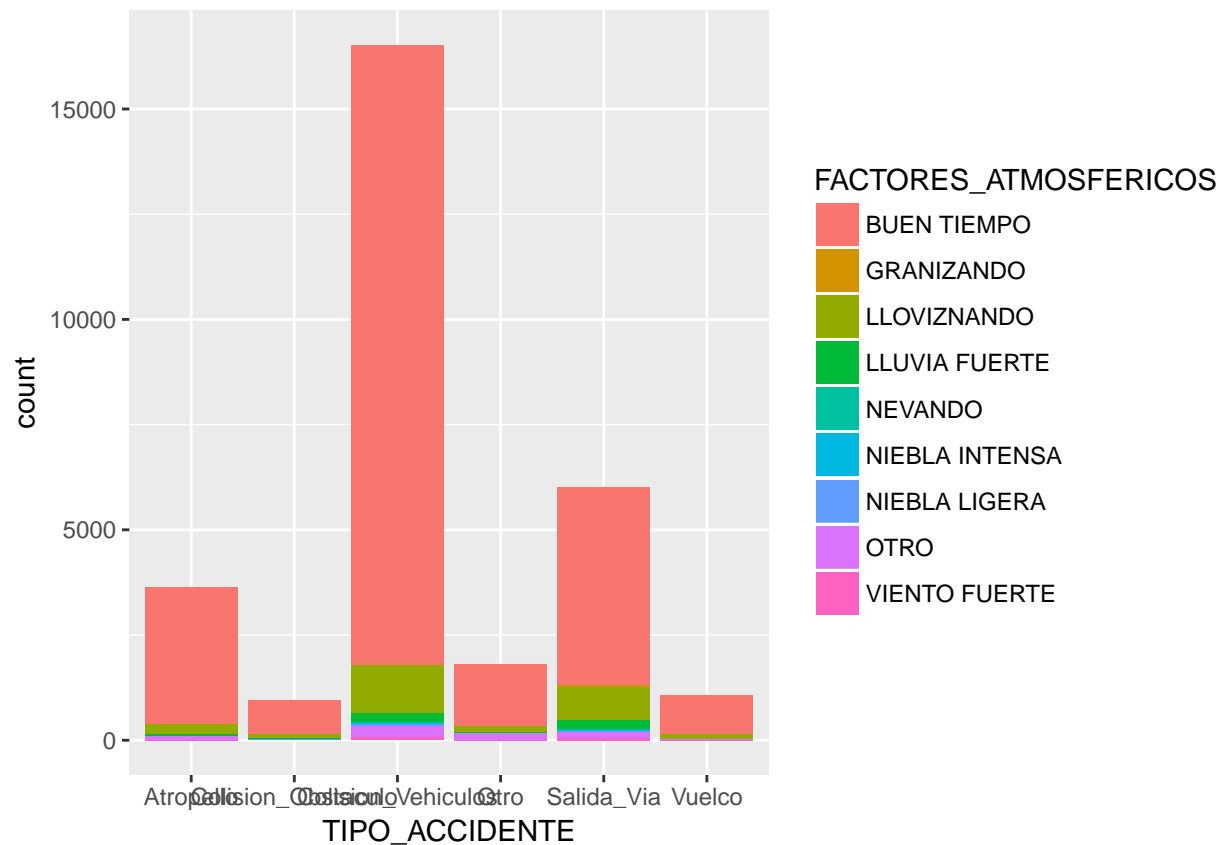
Podemos ver como las zonas predominantes son carretera y zona urbana, pero no parece que esta variable pueda ser influyente a la hora de decir que tipo de accidente se produce por lo que eliminaré esta variable para futuras pruebas.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=RED_CARRETERA))
```



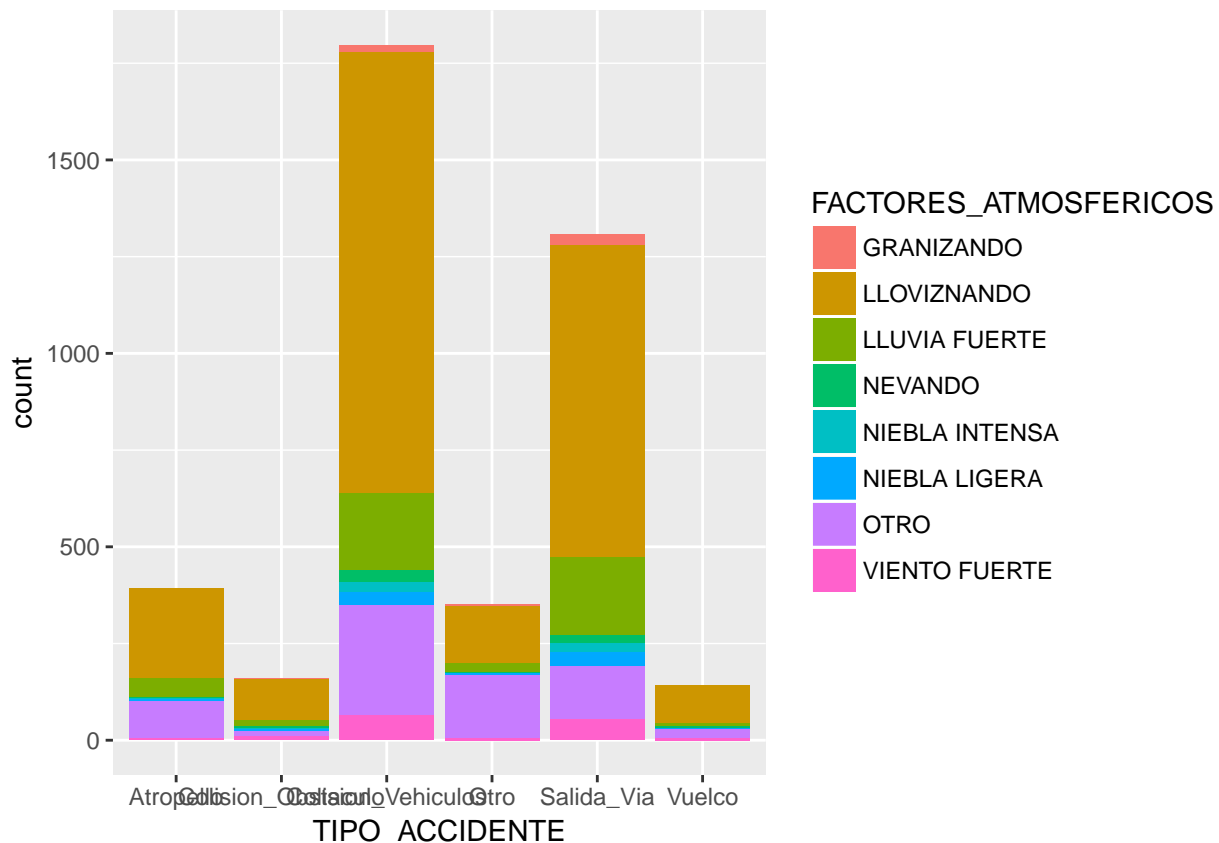
Puede parecer que esta variable no tiene demasiado que ver con la variable que queremos predecir por lo que puede ser que la descartemos.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=FACTORES
```



Por el conocimiento que tenemos, seguramente esta variable no sea demasiado importante para el tipo de accidente. Veamos que le ocurre si eliminamos los elementos que tienen buen tiempo.

```
vector.buen.tiempo <- accidentes.train.sin.variables.1$FACTORES_ATMOSFERICOS == "BUEN TIEMPO"
valores.sin.buen.tiempo <- accidentes.train.sin.variables.1[!vector.buen.tiempo,]
ggplot(data = valores.sin.buen.tiempo) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=FACTORES_ATMOSFERICOS))
```



Pero seguimos viendo que no se puede sacar ninguna conclusión de esta visualización.

3.2 Análisis de variables eliminadas sin valores perdidos

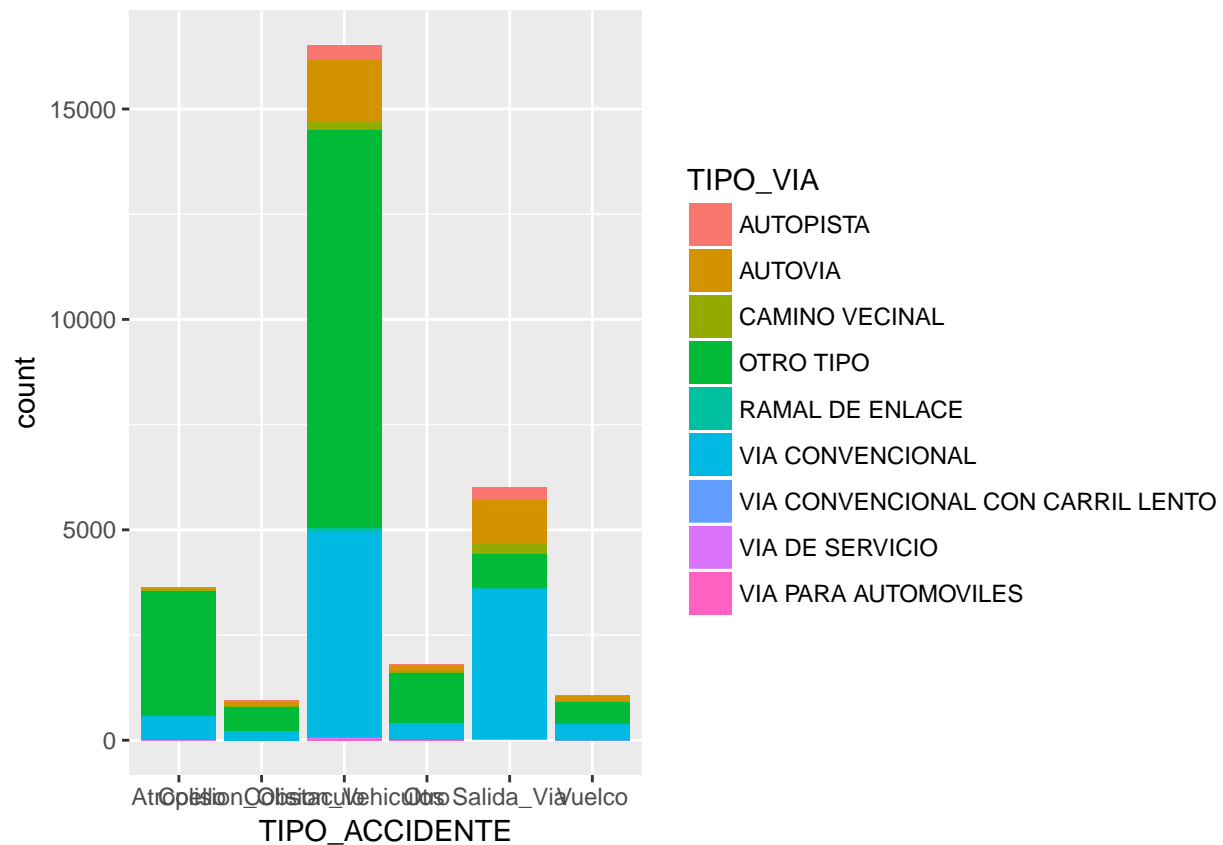
Recordemos las variables que eliminamos sin tener valores perdidos.

```
segundas.variables.eliminadas
```

```
## [1] "ANIO" "MES" "HORA"
## [4] "DIASEMANA" "PROVINCIA" "COMUNIDAD_AUTONOMA"
## [7] "ISLA" "ZONA_AGRUPADA" "TIPO_VIA"
## [10] "TRAZADO_NO_INTERSEC" "TIPO_INTERSEC" "SUPERFICIE_CALZADA"
## [13] "LUMINOSIDAD"
```

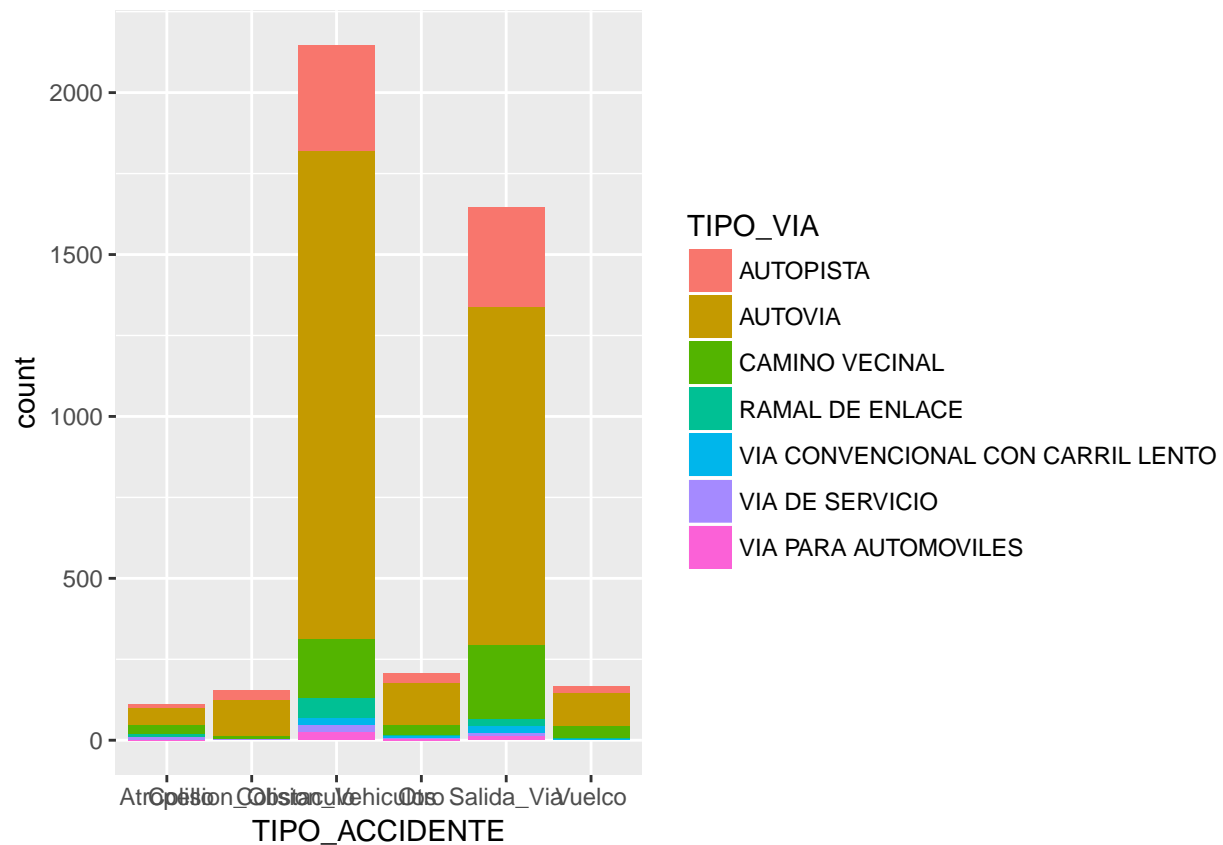
Una de la variables que podrían ser interesantes es TIPO_VIA, TRAZADO_NO_INTERSEC, TIPO_INTERSEC, SUPERFICIE_CALZADA y LUMINOSIDAD. Veamos visualizaciones de estas variables.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TIPO_VIA,
```



Eliminemos las instancias con OTRO TIPO o VIA CONVENCIONAL

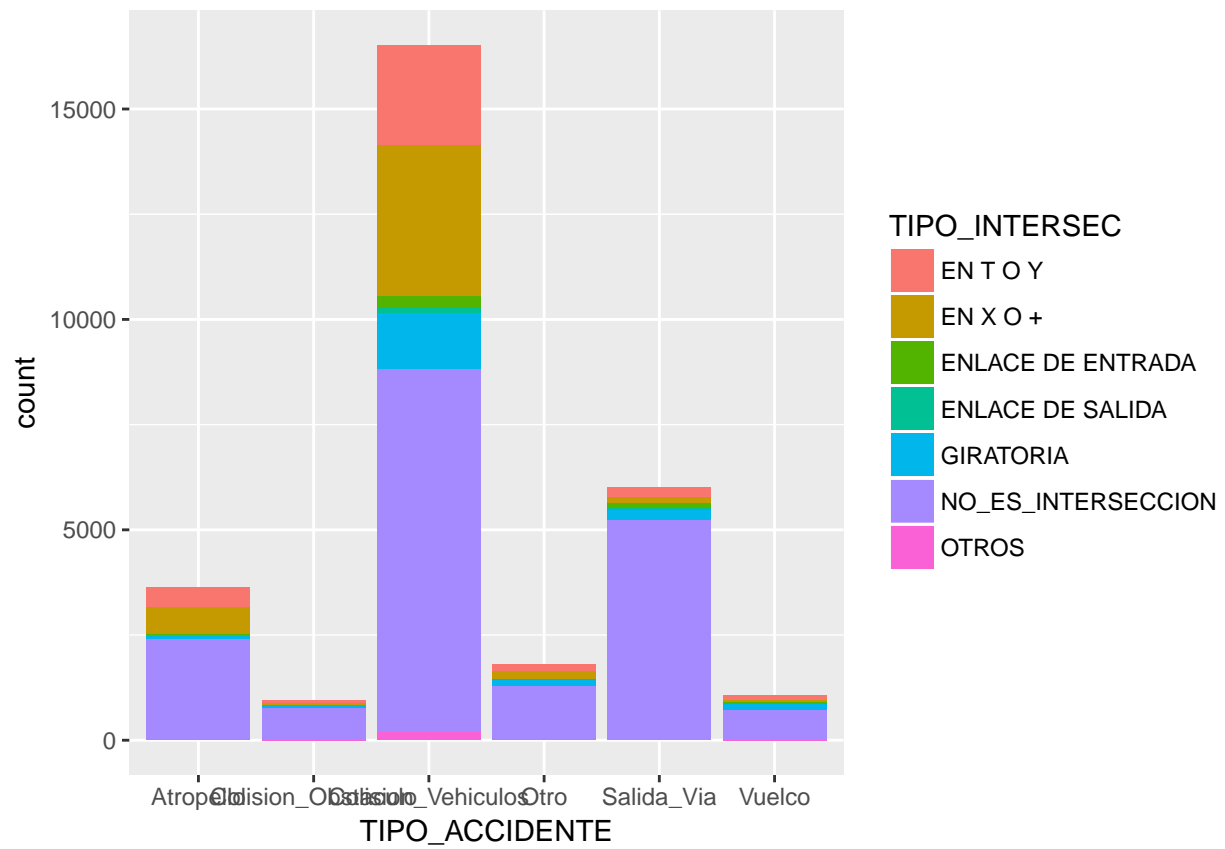
```
vector.sin.otrotipo.y.viaconvencional <- ((accidentes.train.sin.variables.1$TIPO_VIA == "OTRO TIPO") |
valores.sin.otrotipo.y.viaconvencional <- accidentes.train.sin.variables.1[!vector.sin.otrotipo.y.viaconvencional]
ggplot(data = valores.sin.otrotipo.y.viaconvencional) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TIPO_VIA))
```



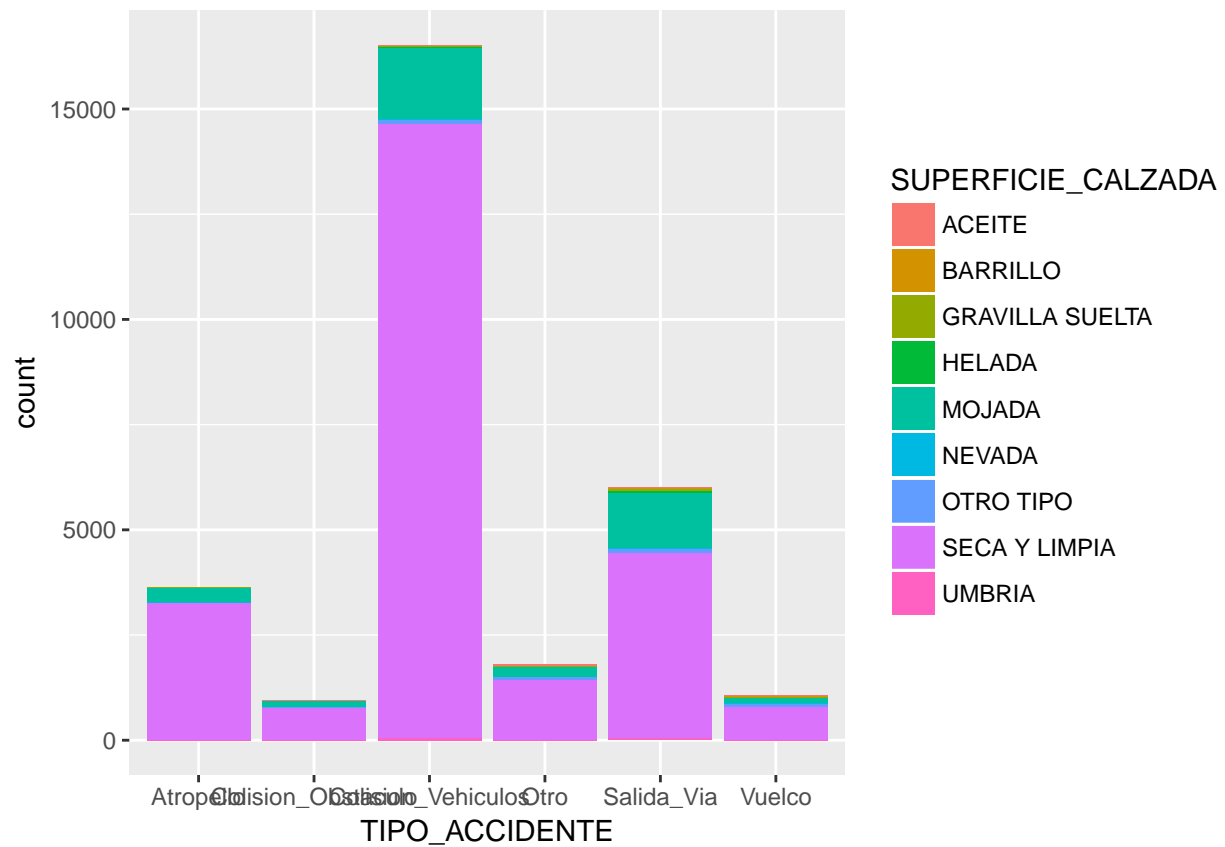
No

se observa que sea una variable demasiada importante.

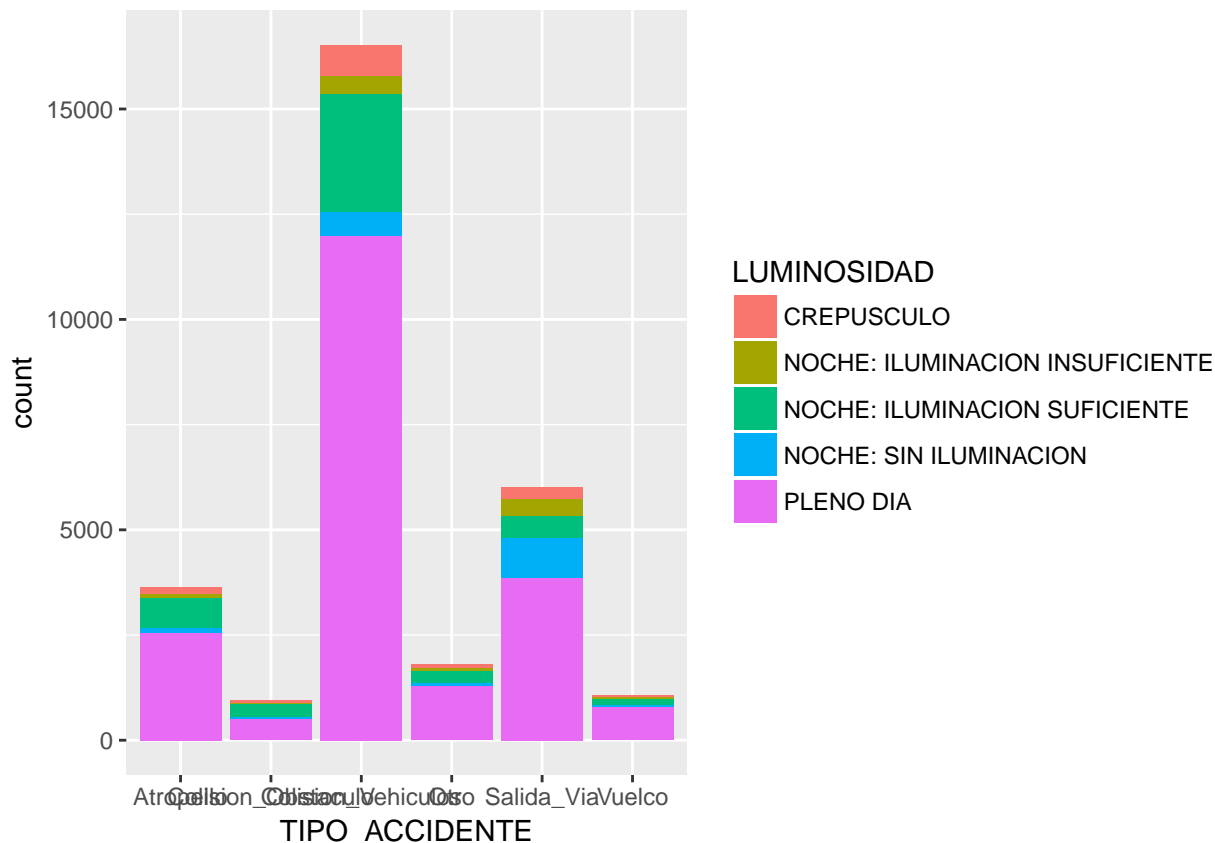
```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TRAZADO))
```

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TIPO_INTERSEC))
```

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=LUMINOS))
```



Por lo que no podemos sacar demasiada información así que no añadiremos ninguna a las que ya estamos usando de momento.

4 Visión preeliminar de los datos

Como anteriormente ya hicimos el summary, no será necesario volver a hacerlo. Lo que si vamos a hacer es un str, para obtener la información de las variables.

```
str(accidentes.train.sin.variables.1)
```

```
## 'data.frame': 30002 obs. of 22 variables:
## $ ANIO : int 2009 2011 2008 2013 2009 2008 2010 2010 2013 2009 ...
## $ MES : Factor w/ 12 levels "Abril","Agosto",...: 8 5 8 10 1 6 6 7 11 10 ...
## $ HORA : Factor w/ 448 levels "0","0,016666667",...: 266 266 136 328 49 411 31 13 ...
## $ DIASEMANA : Factor w/ 7 levels "DOMINGO","JUEVES",...: 7 3 6 7 7 6 4 1 7 6 ...
## $ PROVINCIA : Factor w/ 52 levels "Albacete","Alicante/Alacant",...: 13 39 49 11 2 23 ...
## $ COMUNIDAD_AUTONOMA : Factor w/ 18 levels "Andalucia","Aragon",...: 1 13 11 7 11 1 9 11 14 9 ...
## $ ISLA : Factor w/ 10 levels "FORMENTERA","FUERTEVENTURA",...: 9 9 9 9 9 9 9 9 9 9 ...
## $ TOT_VICTIMAS : int 1 1 1 3 1 2 3 1 1 1 ...
## $ TOT_MUERTOS : int 0 0 1 0 0 1 0 0 0 0 ...
## $ TOT_HERIDOS_GRAVES : int 0 0 0 0 0 1 0 0 0 0 ...
## $ TOT_HERIDOS_LEVES : int 1 1 0 3 1 0 3 1 1 1 ...
## $ TOT_VEHICULOS_IMPLICADOS: int 2 2 1 3 1 1 3 2 1 4 ...
## $ ZONA : Factor w/ 4 levels "CARRETERA","TRAVESIA",...: 4 1 1 4 1 1 4 4 4 4 ...
## $ ZONA_AGRUPADA : Factor w/ 2 levels "VIAS INTERURBANAS",...: 2 1 1 2 1 1 2 2 2 2 ...
## $ RED_CARRETERA : Factor w/ 5 levels "OTRAS TITULARIDADES",...: 4 2 5 4 3 5 4 4 4 4 ...
## $ TIPO_VIA : Factor w/ 9 levels "AUTOPISTA","AUTOVIA",...: 4 6 6 4 1 6 4 4 4 4 ...
```

```
## $ TRAZADO_NO_INTERSEC : Factor w/ 6 levels "CURVA FUERTE CON MARCA Y SIN VELOCIDAD MARCADA",...
## $ TIPO_INTERSEC       : Factor w/ 7 levels "EN T O Y","EN X O +",...: 6 1 6 6 6 6 1 2 6 6 ...
## $ SUPERFICIE_CALZADA  : Factor w/ 9 levels "ACEITE","BARRILLO",...: 8 8 8 5 8 8 8 8 8 ...
## $ LUMINOSIDAD         : Factor w/ 5 levels "CREPUSCULO","NOCHE: ILUMINACION INSUFICIENTE",...: 5
## $ FACTORES_ATMOSFERICOS : Factor w/ 9 levels "BUEN TIEMPO",...: 1 1 1 3 1 1 1 1 1 ...
## $ TIPO_ACCIDENTE      : Factor w/ 6 levels "Atropello","Colision_Obstaculo",...: 3 3 5 3 5 5 3 3
```

Si queremos información más detallada:

```
describe(accidentes.train.sin.variables.2[1])
```

```
## accidentes.train.sin.variables.2[1]
##
## 1 Variables      30002 Observations
## -----
## TOT_VICTIMAS
##      n missing distinct      Info      Mean      Gmd      .05      .10
## 30002      0      17    0.609    1.429    0.6909      1      1
##    .25    .50    .75    .90    .95
##      1      1      2      2      3
##
## Value      1      2      3      4      5      6      7      8      9     10
## Frequency 21826  5503  1540   681   248   105   43   25   13    8
## Proportion 0.727 0.183 0.051 0.023 0.008 0.003 0.001 0.001 0.000 0.000
##
## Value      11      12      13      15      17      18      19
## Frequency      3      1      2      1      1      1      1
## Proportion 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## -----
```

Esto lo podemos hacer con las variables que veamos oportunas. Otra forma de ver más información es:

```
basicStats(accidentes.train.sin.variables.2[1])
```

```
##      TOT_VICTIMAS
## nobs      30002.000000
## NAs       0.000000
## Minimum   1.000000
## Maximum   19.000000
## 1. Quartile 1.000000
## 3. Quartile 2.000000
## Mean       1.429371
## Median     1.000000
## Sum        42884.000000
## SE Mean    0.005258
## LCL Mean   1.419066
## UCL Mean   1.439677
## Variance   0.829334
## Stdev      0.910678
## Skewness   3.817690
## Kurtosis   27.886723
```

5 Imputación de valores perdidos

Vamos a usar uso del paquete mice para imputar los datos.

5.1 Imputación de variables

Veamos que variables teníamos con valores perdidos.

```
summary(accidentes.train.variables.eliminadas)
```

```
##      CARRETERA
## A-7      : 294
## A-2      : 278
## AP-7     : 229
## N-340    : 229
## A-4      : 184
## (Other):12098
## NA's     :16690
##
##                                ACOND_CALZADA
## CARRIL CENTRAL DE ESPERA      : 193
## NADA ESPECIAL                 : 4645
## OTRO TIPO                     : 791
## PASO PARA PEATONES O ISLETAS EN CENTRO DE VIA PRINCIPAL: 397
## RAQUETA DE GIRO IZQUIERDA    : 109
## SOLO ISLETAS O PASO PARA PEATONES : 168
## NA's                         :23699
##
##          PRIORIDAD          VISIBILIDAD_RESTRINGIDA
## NINGUNA (SOLO NORMA) :13495 SIN RESTRICCION      :16982
## SEMAFORO             : 1778 CONFIGURACION DEL TERRENO: 989
## SEÑAL DE STOP        : 1750 OTRA_CAUSA          : 491
## SOLO MARCAS VIALES   : 1659 FACTORES ATMOSFERICOS : 374
## SEÑAL DE CEDA EL PASO: 1629 EDIFICIOS           : 229
## (Other)              : 1569 (Other)            : 252
## NA's                 : 8122 NA's              :10685
##
##          OTRA_CIRCUNSTANCIA          ACERAS          DENSIDAD_CIRCULACION
## NINGUNA      :24967 NO HAY ACERA:21416 CONGESTIONADA: 308
## OTRA         : 942 SI HAY ACERA: 5437 DENSA      : 1479
## OBRAS        : 263 NA's           : 3149 FLUIDA     :17505
## FUERTE DESCENSO : 227 NA's           :10710
## CAMBIO DE RASANTE: 100
## (Other)       : 264
## NA's         : 3239
##
##          MEDIDAS_ESPECIALES
## CARRIL REVERSIBLE : 17
## HABILITACION ARCEN: 8
## NINGUNA MEDIDA    :21024
## OTRA MEDIDA       : 278
## NA's              : 8675
##
##
```

Vemos que dos de estas variables que podrían ser más interesantes son visibilidad restringida y prioridad, por lo que vamos a proceder a imputar sus valores perdidos.

```
accidentes.train.a.imputar <- cbind(accidentes.train.sin.variables.2, accidentes.train.variables.eliminadas)
accidentes.test.a.imputar <- cbind(accidentes.test.sin.variables.2, accidentes.test.variables.eliminadas)
set.seed(1234)
train.imputados.incompletos <- mice::mice(accidentes.train.a.imputar, m=1, method="pmm")
```

```
##
```

```
## iter imp variable
## 1 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA

train.imputados <- mice::complete(train.imputados.incompletos)
test.imputados.incompletos <- mice::mice(accidentes.test.a.imputar, m=5, method="pmm")

##
## iter imp variable
## 1 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA

test.imputados <- mice::complete(test.imputados.incompletos)
```

5.2 Prueba del modelo con imputación de valores perdidos

Hagamos por lo tanto una prueba de como afecta la imputación de valores perdidos.

```
set.seed(1234)
ct3 <- ctree(TIPO_ACCIDENTE ~., train.imputados)
testPred3 <- predict(ct3, newdata = test.imputados)
```

Por lo que ya tenemos el conjunto de test predicho. Además el árbol creado tendría la siguiente estructura:

```
#ct3
```

Vamos a escribir la salida del modelo para ver su puntuación en Kaggel.

```
salida.tercer.modelo <- as.matrix(testPred3)
salida.tercer.modelo <- cbind(c(1:(dim(salida.tercer.modelo)[1])), salida.tercer.modelo)
```

```
colnames(salida.tercer.modelo) <- c("Id","Prediction")
write.table(salida.tercer.modelo,file="predicciones/TerceraPrediccion.txt",sep="," ,quote = F,row.names =
```

El resultado de este modelo para la competición de Kaggel, subido el 19/02/2017 a las 17:42, con un total de 14 personas entregadas, se ha quedado en la posición 9 con una puntuación del 0.81753. Bajando muy poco con respecto a la anterior puntuación.

#	Δ5d	Team Name	Score 	Entries	Last Submission UTC (Best ~ Last Submission)
1	new	Anabel Gómez	0.83175	12	Sun, 19 Feb 2017 13:06:40 (-2.9d)
2	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
3	new	Jonathan Espinosa	0.82780	12	Sun, 19 Feb 2017 11:41:59
4	↓2	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
5	↓1	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
6	↓3	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
7	new	WhiteShadow	0.82345	6	Sat, 18 Feb 2017 14:23:36 (-17.9h)
8	new	Jorge Jimena	0.82059	4	Sun, 19 Feb 2017 16:12:15 (-0.2h)
9	↓3	PacoPollos	0.81891	3	Sun, 19 Feb 2017 16:41:50 (-47.9h)
Your Best Entry ↑ Your submission scored 0.81753 , which is not an improvement of your best score. Keep trying!					
10	↓5	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
11	↓4	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
12	↓4	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
13	new	Xisco Fauli	0.48735	2	Wed, 15 Feb 2017 23:16:45
14	↓5	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 3: Tercera puntuación obtenida en Kaggel

6 Detección de anomalías

Veamos como detectar valores anómalos en nuestros datos.

6.1 Uso del paquete outliers

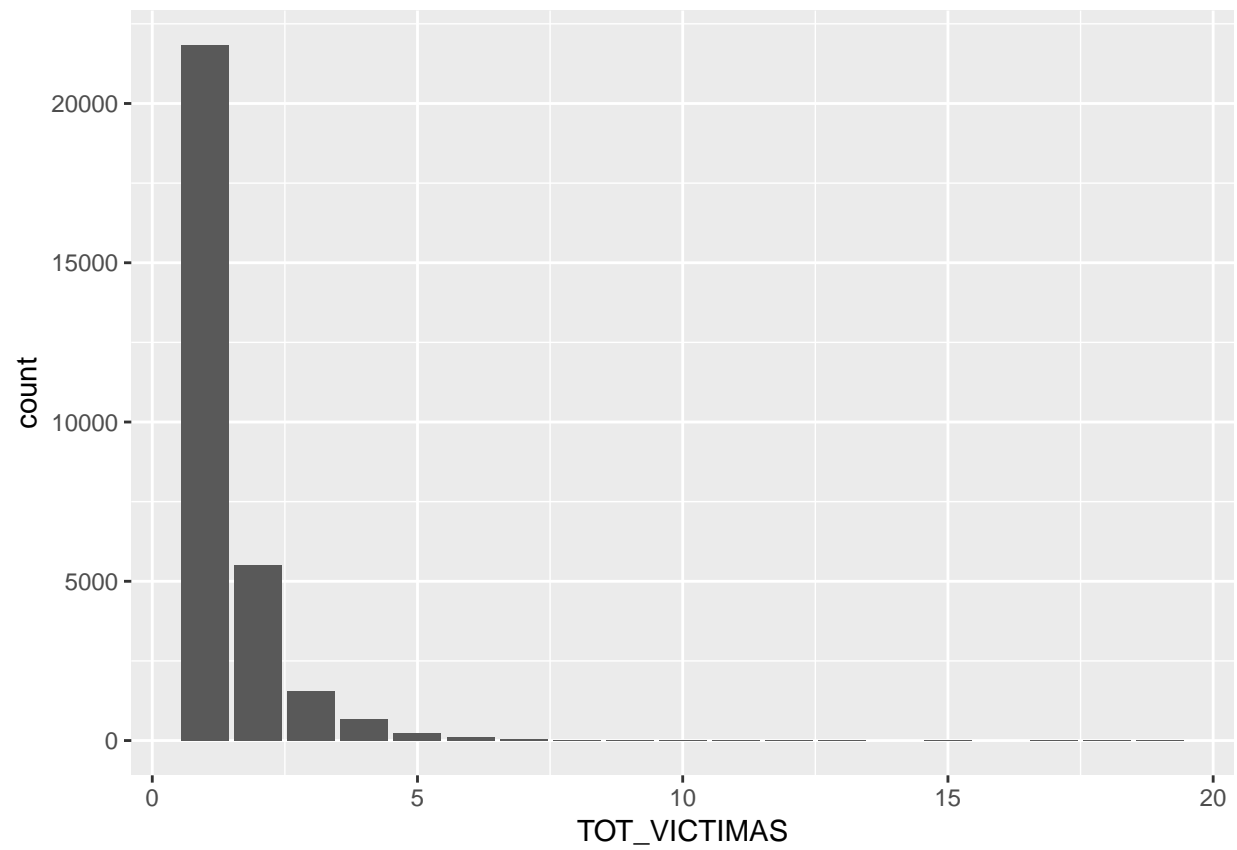
Veamos si tenemos valores perdidos en nuestros datos, solo con valores que no son discretas.

```
valores.anomalos <- outliers::outlier(train.imputados[,1:5])
print(valores.anomalos)
```

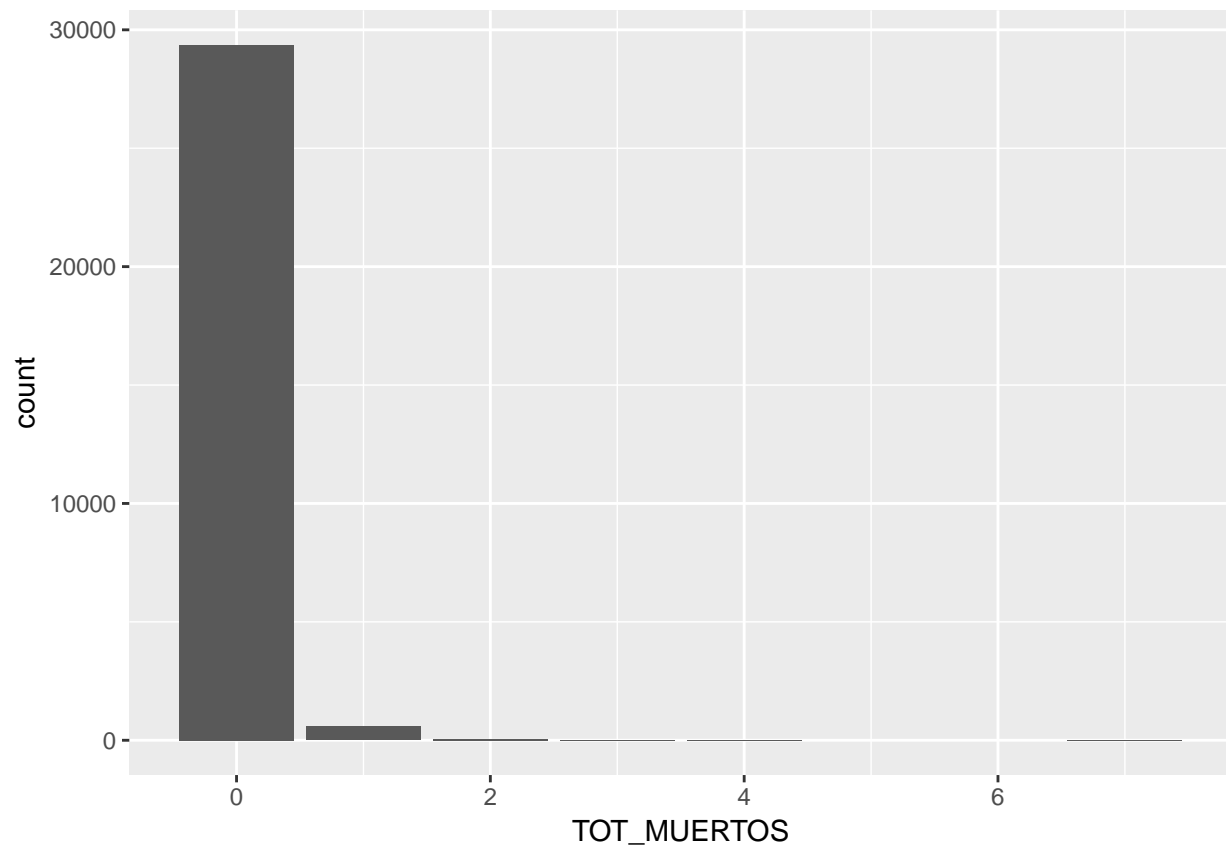
```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##                  19                      7                      9
```

```
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS  
##                18                21
```

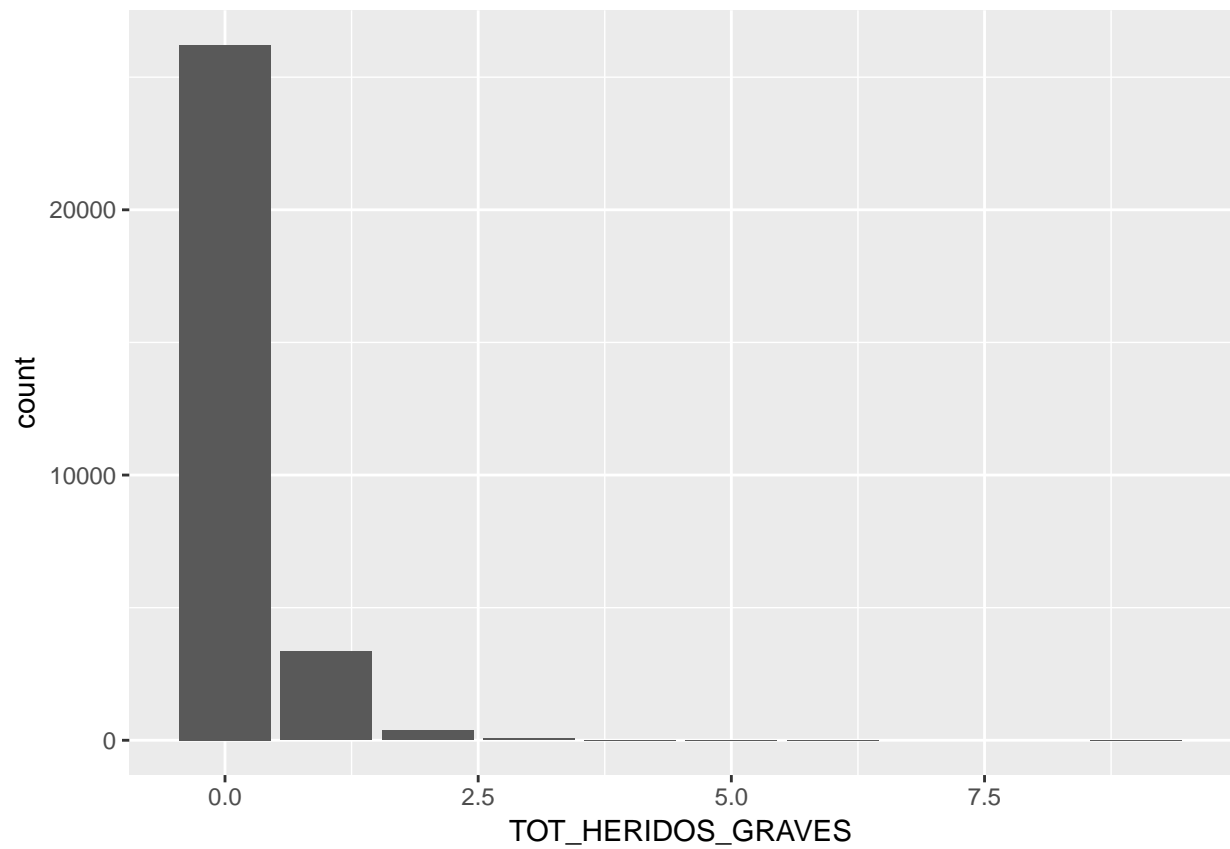
```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_VICTIMAS))
```



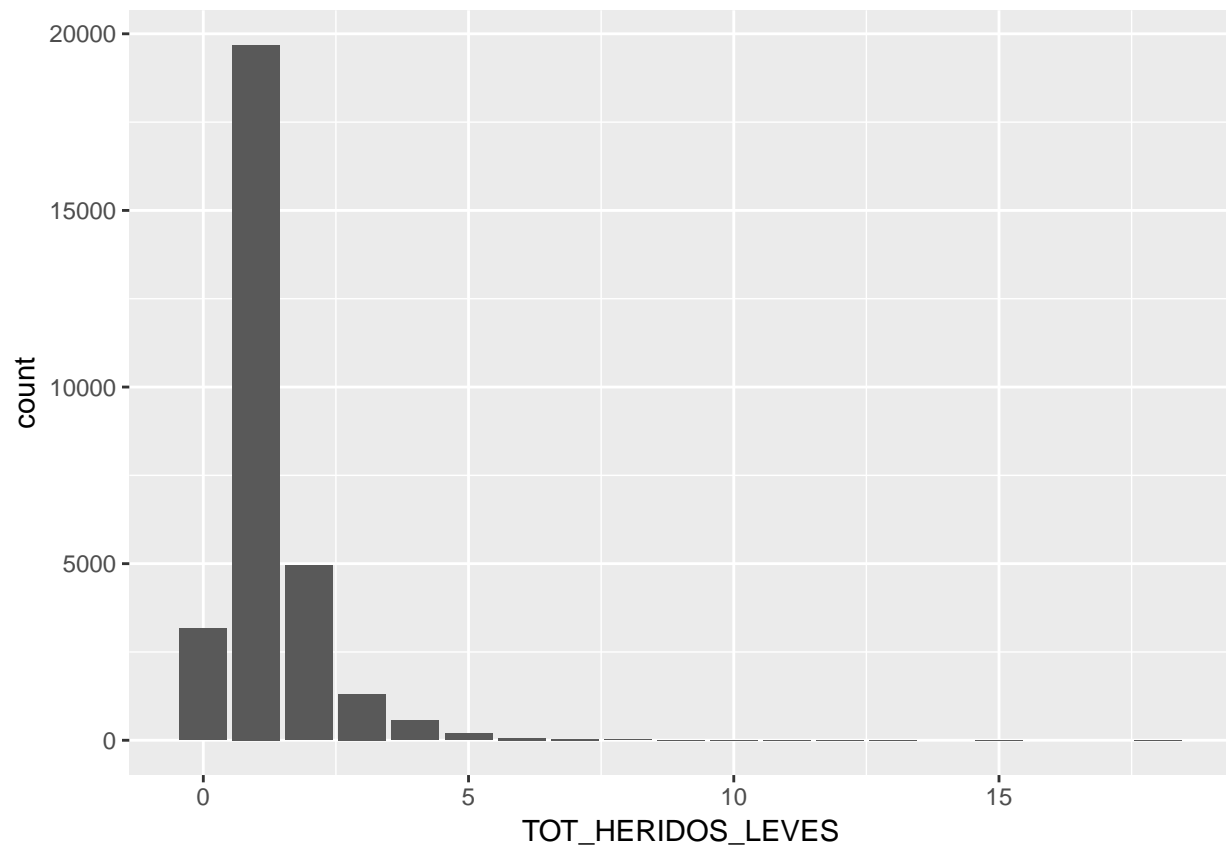
```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_MUERTOS))
```



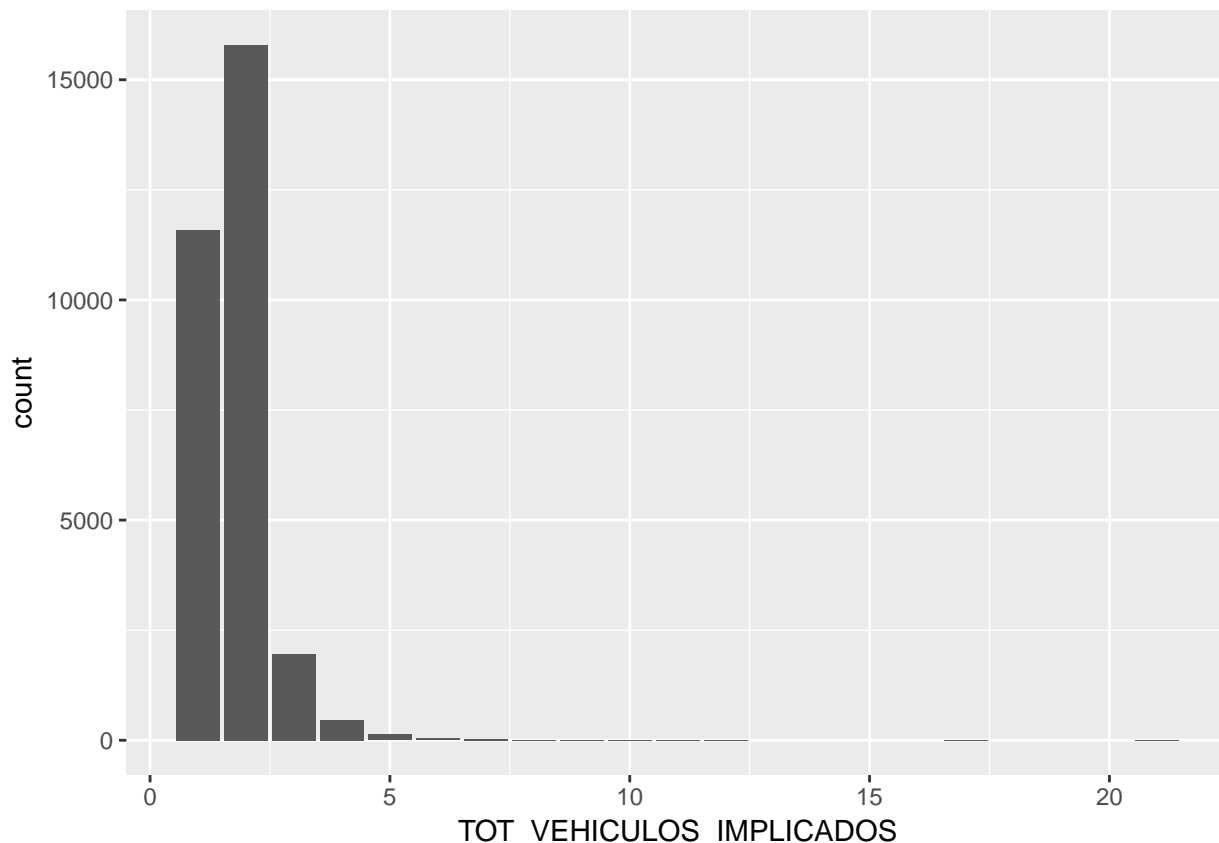
```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_HERIDOS_GRAVES))
```

```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_HERIDOS_GRAVES))
```



```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_VEHICULOS_IMPLICADOS))
```



Viendo que en cada variable tenemos distintos valores anómalos como sería el valor 19 en TOT_VICTIMAS.

6.2 Paquete mvoutlier

Voy a intentar usar el paquete mvoutlier.

```
require(mvoutlier)
```

```
## Loading required package: mvoutlier
```

```
## Warning: package 'mvoutlier' was built under R version 3.3.2
```

```
## Loading required package: sgeostat
```

```
## sROC 0.1-2 loaded
```

```
#resultado.búsqueda.anomalias <- uni.plot(train.imputados[1:200,1:2])
```

Como se puede ver, se ha obtenido un error el cual no he podido solucionar.

6.3 Eliminación de valores anómalos

En función de lo obtenido con el paquete outlier, voy a intentar realizar algo con este paquete para ver que tal se comporta nuestro dataset.

```
valores.anomalos.train <- outliers::outlier(train.imputados[,1:5])
valores.anomalos.test <- outliers::outlier(test.imputados[,1:5])
print(valores.anomalos.train)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##              19              7              9
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##              18              21
```

```
print(valores.anomalos.test)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##              10              5              5
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##              10              11
```

Veamos, por ejemplo, para la variable TOT_VICTIMAS, cuantas instancias cumplen tener mas de 19 victimas o de 10.

```
vector.con.victimas.19 <- train.imputados$TOT_VICTIMAS >= 19
sum(vector.con.victimas.19)
```

```
## [1] 1
```

```
vector.con.victimas.18 <- train.imputados$TOT_VICTIMAS >= 18
sum(vector.con.victimas.18)
```

```
## [1] 2
```

```
vector.con.victimas.17 <- train.imputados$TOT_VICTIMAS >= 17
sum(vector.con.victimas.17)
```

```
## [1] 3
```

```
vector.con.victimas.10 <- train.imputados$TOT_VICTIMAS >= 10
sum(vector.con.victimas.10)
```

```
## [1] 18
```

```
valores.con.victimas.10 <- train.imputados[vector.con.victimas.10,]
valores.con.victimas.10$TIPO_ACCIDENTE
```

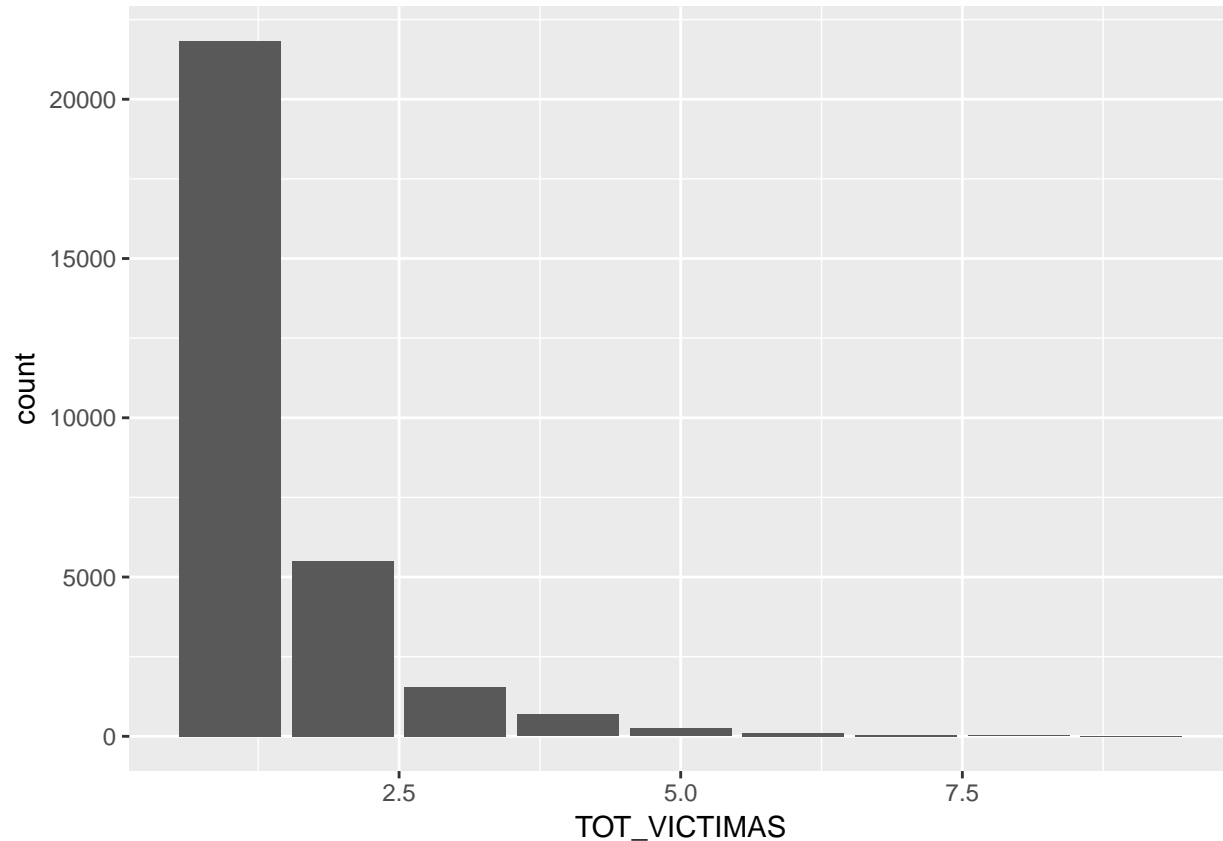
```
## [1] Salida_Via          Colision_Vehiculos Colision_Vehiculos
## [4] Colision_Vehiculos Colision_Vehiculos Colision_Vehiculos
## [7] Salida_Via          Colision_Vehiculos Salida_Via
## [10] Colision_Vehiculos Colision_Vehiculos Colision_Vehiculos
## [13] Salida_Via          Otro              Colision_Vehiculos
## [16] Colision_Vehiculos Colision_Vehiculos Colision_Vehiculos
## attr(,"contrasts")
##              2 3 4 5 6
## Atropello      0 0 0 0 0
## Colision_Obstaculo 1 0 0 0 0
## Colision_Vehiculos 0 1 0 0 0
## Otro           0 0 1 0 0
## Salida_Via     0 0 0 1 0
## Vuelco         0 0 0 0 1
## 6 Levels: Atropello Colision_Obstaculo Colision_Vehiculos ... Vuelco
```

Vemos que no son demasiados datos, ya que en total son 18 instancias, por lo que vamos a probar a eliminarlas a ver el comportamiento del paquete outlier de nuevo.

```
train.sin.outliers <- train.imputados[!vector.con.victimas.10,]
valores.anomalos.sin.victimas.10 <- outliers::outlier(train.sin.outliers[,1:5])
print(valores.anomalos.sin.victimas.10)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##              9              7              6
##    TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##              9              17
```

```
ggplot(data = train.sin.outliers) + geom_bar(mapping = aes(x=TOT_VICTIMAS))
```



Vamos a probar a eliminar algunas instancias, con los criterios de otras variables.

```
vector.con.muertos.7 <- train.sin.outliers$TOT_MUERTOS >= 7
sum(vector.con.muertos.7)
```

```
## [1] 1
```

```
vector.con.muertos.6 <- train.sin.outliers$TOT_MUERTOS >= 6
sum(vector.con.muertos.6)
```

```
## [1] 1
```

```
vector.con.muertos.5 <- train.sin.outliers$TOT_MUERTOS >= 5
sum(vector.con.muertos.5)
```

```
## [1] 1
```

```
vector.con.muertos.4 <- train.sin.outliers$TOT_MUERTOS >= 4
sum(vector.con.muertos.4)
```

```
## [1] 6
```

```
train.sin.outliers <- train.sin.outliers[!vector.con.muertos.4,]
valores.anomalos.sin.muertos.4 <- outliers::outlier(train.sin.outliers[,1:5])
print(valores.anomalos.sin.muertos.4)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##                9                3                6
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##                9                17
```

Vamos a realizarlo más rápidamente

```
vector.con.anomalias <- ((train.sin.outliers$TOT_HERIDOS_GRAVES >= 6) | (train.sin.outliers$TOT_HERIDOS_LEVES >= 6))
sum(vector.con.anomalias)
```

```
## [1] 10
```

```
vector.con.anomalias <- ((train.sin.outliers$TOT_HERIDOS_GRAVES >= 5) | (train.sin.outliers$TOT_HERIDOS_LEVES >= 5))
sum(vector.con.anomalias)
```

```
## [1] 31
```

```
train.sin.outliers <- train.sin.outliers[!vector.con.anomalias,]
```

Pero, que pasaría si eliminamos en función de las anomalías que nos marca el test:

```
print(valores.anomalos.test)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##                10                5                5
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##                10                11
```

```
vector.con.anomalias <- ((train.imputados$TOT_HERIDOS_GRAVES > 5) | (train.imputados$TOT_HERIDOS_LEVES > 5))
sum(vector.con.anomalias)
```

```
## [1] 14
```

En total eliminaríamos 14 instancias. Vamos a comprobarlo:

```
train.sin.outliers <- train.imputados[!vector.con.anomalias,]
valores.anomalos.train <- outliers::outlier(train.sin.outliers[,1:5])
print(valores.anomalos.train)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##                10                4                5
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##                10                11
```

```
print(valores.anomalos.test)
```

```
##          TOT_VICTIMAS          TOT_MUERTOS          TOT_HERIDOS_GRAVES
##                10                5                5
##      TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##                10                11
```

6.4 Prueba del modelo con imputación de valores perdidos

Hagamos por lo tanto una prueba de como afecta la imputación de valores perdidos.

```
set.seed(1234)
ct4 <- ctree(TIPO_ACCIDENTE ~., train.sin.outliers)
testPred4 <- predict(ct4, newdata = test.imputados)
```

Por lo que ya tenemos el conjunto de test predicho. Además el árbol creado tendría la siguiente estructura:

```
#ct4
```

Vamos a escribir la salida del modelo para ver su puntuación en Kaggel.

```
salida.modelo.4 <- as.matrix(testPred4)
salida.modelo.4 <- cbind(c(1:(dim(salida.modelo.4)[1])), salida.modelo.4)
colnames(salida.modelo.4) <- c("Id", "Prediction")
write.table(salida.modelo.4, file="predicciones/Prediccion4.txt", sep="," , quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 19/02/2017 a las 20:12, con un total de 16 personas entregadas, se ha quedado en la posición 10 con una puntuación del 0.81753. Bajando muy poco con respecto a la anterior puntuación.

#	Δ5d	Team Name	Score ?	Entries	Last Submission UTC (Best - Last Submission)
1	new	Anabel Gómez	0.83175	12	Sun, 19 Feb 2017 13:06:40 (-2.9d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	new	Jonathan Espinosa	0.82780	13	Sun, 19 Feb 2017 18:42:10 (-7h)
5	↓3	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
6	↓2	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
7	↓4	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
8	new	WhiteShadow	0.82345	6	Sat, 18 Feb 2017 14:23:36 (-17.9h)
9	new	Jorge Jimena	0.82059	4	Sun, 19 Feb 2017 16:12:15 (-0.2h)
10	↓4	PacoPollos	0.81891	4	Sun, 19 Feb 2017 19:11:28 (-2.1d)
Your Best Entry ↑ Your submission scored 0.81753 , which is not an improvement of your best score. Keep trying!					
11	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
12	↓7	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
13	↓6	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
14	↓6	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
15	new	Xisco Fauli	0.48735	2	Wed, 15 Feb 2017 23:16:45
16	↓7	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 4: Cuarta puntuación obtenida en Kaggel

7 Transformación de los datos

Tal y como se vio en el guión de prácticas en el punto 7, vamos a aplicar la transformación para ver que tal nos funciona.

7.1 Transformando los datos

Vamos a aplicar centrado y escalado sobre el conjunto de datos con los valores ya imputados, para las variables que se consideran continuas.

```
valores.preprocesados <- caret::preProcess(train.sin.outliers[,1:5],method=c("center","scale"))
valores.transofrmados <- predict(valores.preprocesados,train.sin.outliers[,1:5])
train.transformado <- cbind(valores.transofrmados,train.sin.outliers[,6:11])
valores.preprocesados.test <- caret::preProcess(test.imputados[,1:5],method=c("center","scale"))
valores.transofrmados.test <- predict(valores.preprocesados.test,test.imputados[,1:5])
test.transformado <- cbind(valores.transofrmados.test,test.imputados[,6:10])
```

7.2 Prueba del modelo con transformación de los datos

Hagamos por lo tanto una prueba de como afecta la transformación de los datos.

```
set.seed(1234)
ct5 <- ctree(TIPO_ACCIDENTE ~., train.transformado)
testPred5 <- predict(ct5, newdata = test.transformado)
```

Por lo que ya tenemos el conjunto de test predicho. Además el árbol creado tendría la siguiente estructura:

```
#ct5
```

Vamos a escribir la salida del modelo para ver su puntuación en Kaggel.

```
salida.modelo.5 <- as.matrix(testPred5)
salida.modelo.5 <- cbind(c(1:(dim(salida.modelo.5)[1])), salida.modelo.5)
colnames(salida.modelo.5) <- c("Id","Prediction")
write.table(salida.modelo.5,file="predicciones/Prediccion5.txt",sep="," ,quote = F,row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 20/02/2017 a las 13:15, con un total de 18 personas entregadas, se ha quedado en la posición 12 con una puntuación del 0.55147. Bajando mucho con respecto a la anterior puntuación, por lo que esta transformación no la tendremos en cuenta.

8 Discretización

Para este conjunto de datos no se realiza discretización ya que no tenemos variables continuas como para poder discretizarlas.

9 Selección de características

Para este apartado comenzaremos con los dataset originales.

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
```


#	$\Delta 6d$	Team Name	Score ?	Entries	Last Submission UTC (Best – Last Submission)
1	new	Anabel Gómez	0.83175	15	Mon, 20 Feb 2017 07:44:42 (-3.7d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	new	Jonathan Espinosa	0.82859	14	Sun, 19 Feb 2017 19:39:14
5	new	Xisco Fauli	0.82810	10	Mon, 20 Feb 2017 10:50:53 (-1.3h)
6	new	ManuelMontero	0.82582	3	Sun, 19 Feb 2017 20:10:00
7	↓5	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
8	↓5	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
9	new	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
10	new	WhiteShadow	0.82345	6	Sat, 18 Feb 2017 14:23:36 (-17.9h)
11	new	Jorge Jimena	0.82306	6	Sun, 19 Feb 2017 20:48:53
12	↓7	PacoPollos	0.81891	5	Mon, 20 Feb 2017 12:15:12 (-2.8d)
Your Best Entry ↑ Your submission scored 0.55147 , which is not an improvement of your best score. Keep trying!					
13	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
14	new	Salva Moreno	0.81891	2	Mon, 20 Feb 2017 12:00:33
15	↓11	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
16	↓10	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
17	↓10	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
18	↓10	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 5: Quinta puntuación obtenida en Kaggel

9.1 Paquete FSelector

9.1.1 Aproximación filter: chi.squared

Determina los pesos de los atributos discretos usando el test de independencia chi-cuadrado (con respecto a la variable clase). Calculamos los pesos de los atributos: la medida devuelta indica el nivel de dependencia de cada atributo frente a la variable clase

```
set.seed(1234)
pesos <- FSelector::chi.squared(TIPO_ACCIDENTE~,train.original)
pesos
```

```
##                attr_importance
## ANIO                0.00000000
## MES                 0.03094819
## HORA                0.15532002
## DIASEMANA           0.06981337
## PROVINCIA           0.14327070
## COMUNIDAD_AUTONOMA  0.12198994
## ISLA                0.02436129
## TOT_VICTIMAS         0.09669636
## TOT_MUERTOS          0.06428765
## TOT_HERIDOS_GRAVES  0.08816985
## TOT_HERIDOS_LEVES   0.13237988
## TOT_VEHICULOS_IMPLICADOS 0.63503097
## ZONA                0.26553819
## ZONA_AGRUPADA        0.45894923
## CARRETERA           0.52879460
## RED_CARRETERA        0.19748117
## TIPO_VIA            0.18875170
## TRAZADO_NO_INTERSEC 0.19181245
## TIPO_INTERSEC        0.14228134
## ACOND_CALZADA        0.09438668
## PRIORIDAD           0.22060851
## SUPERFICIE_CALZADA  0.11239155
## LUMINOSIDAD          0.12226652
## FACTORES_ATMOSFERICOS 0.08055707
## VISIBILIDAD_RESTRINGIDA 0.09419773
## OTRA_CIRCUNSTANCIA   0.06053977
## ACERAS              0.22765102
## DENSIDAD_CIRCULACION 0.12681797
## MEDIDAS_ESPECIALES  0.04702684
```

Vamos a seleccionar los 7 mejores

```
subset <- FSelector::cutoff.k(pesos, 7)
las.7.mas.importantes.chi.squared <- as.simple.formula(subset, "TIPO_ACCIDENTE")
las.7.mas.importantes.chi.squared
```

```
## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + CARRETERA + ZONA_AGRUPADA +
##      ZONA + ACERAS + PRIORIDAD + RED_CARRETERA
## <environment: 0x7f9b7002c348>
```

Por lo que vamos a montar un modelo con estas variables

```
train.filter.chi.squared <- train.original[,c("TOT_VEHICULOS_IMPLICADOS", "CARRETERA", "ZONA_AGRUPADA", "ZONA", "ACERAS", "PRIORIDAD", "RED_CARRETERA")]
test.filter.chi.squared <- test.original[,c("TOT_VEHICULOS_IMPLICADOS", "CARRETERA", "ZONA_AGRUPADA", "ZONA", "ACERAS", "PRIORIDAD", "RED_CARRETERA")]
```

```
summary(train.filter.chi.squared)
```

```
## TOT_VEHICULOS_IMPLICADOS CARRETERA ZONA_AGRUPADA
## Min. : 1.000 A-7 : 294 VIAS INTERURBANAS:13335
## 1st Qu.: 1.000 A-2 : 278 VIAS URBANAS :16667
## Median : 2.000 AP-7 : 229
## Mean : 1.738 N-340 : 229
## 3rd Qu.: 2.000 A-4 : 184
## Max. :21.000 (Other):12098
## NA's :16690
## ZONA ACERAS PRIORIDAD
## CARRETERA :13278 NO HAY ACERA:21416 NINGUNA (SOLO NORMA) :13495
## TRAVESIA : 241 SI HAY ACERA: 5437 SEMAFORO : 1778
## VARIANTE : 57 NA's : 3149 SEÑAL DE STOP : 1750
## ZONA URBANA:16426 SOLO MARCAS VIALES : 1659
## SEÑAL DE CEDA EL PASO: 1629
## (Other) : 1569
## NA's : 8122
## RED_CARRETERA
## OTRAS TITULARIDADES : 318
## TITULARIDAD AUTONOMICA : 3890
## TITULARIDAD ESTATAL : 4021
## TITULARIDAD MUNICIPAL :19077
## TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
## TIPO_ACCIDENTE
## Atropello : 3642
## Colision_Obstaculo: 952
## Colision_Vehiculos:16520
## Otro : 1807
## Salida_Via : 6013
## Vuelco : 1068
##
```

Vemos que la variable CARRETERA tiene un alto número de valores perdidos por lo que la vamos a descartar, a pesar de que la selección de características nos ha dicho que es importante.

```
train.filter.chi.squared["CARRETERA"] <- NULL
test.filter.chi.squared["CARRETERA"] <- NULL
```

9.1.2 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct6 <- ctree(TIPO_ACCIDENTE ~., train.filter.chi.squared)
testPred6 <- predict(ct6, newdata = test.filter.chi.squared)
```

Por lo que ya tenemos el conjunto de test predicho. Además el árbol creado tendría la siguiente estructura:

```
#ct6
```

Vamos a escribir la salida del modelo para ver su puntuación en Kaggel.

```
salida.modelo.6 <- as.matrix(testPred6)
salida.modelo.6 <- cbind(c(1:(dim(salida.modelo.6)[1])), salida.modelo.6)
colnames(salida.modelo.6) <- c("Id", "Prediction")
write.table(salida.modelo.6, file="predicciones/Prediccion6.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 22/02/2017 a las 13:20, con un total de 21 personas entregadas, se ha quedado en la posición 13 con una puntuación del 0.82089. Mejorando a la que ya se tenía anteriormente, por lo que vemos que esta selección de características ha funcionado correctamente.

#	Δ6d	Team Name	Score ?	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	—	RubenSanchez	0.82889	12	Tue, 21 Feb 2017 22:52:29
5	↑1	Jonathan Espinosa	0.82859	14	Sun, 19 Feb 2017 19:39:14
6	↑6	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
7	new	Jorge Jimena	0.82662	13	Wed, 22 Feb 2017 11:58:13 (-33.8h)
8	↓1	WhiteShadow	0.82632	9	Wed, 22 Feb 2017 12:02:54 (-20.5h)
9	new	ManuelMontero	0.82582	11	Wed, 22 Feb 2017 11:56:12 (-2.7d)
10	↓5	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
11	new	Salva Moreno	0.82573	10	Wed, 22 Feb 2017 01:48:56 (-25.9h)
12	↓9	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
13	↓4	PacoPollos	0.82089	6	Wed, 22 Feb 2017 12:20:01
<p>Your Best Entry ↑ You improved on your best score by 0.00198.</p> <p>You just moved up 1 position on the leaderboard. Tweet this!</p>					
14	new	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
15	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
16	↓8	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
17	new	Mauricio Orellana	0.73246	1	Wed, 22 Feb 2017 02:42:56
18	↓8	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
19	↓8	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
20	new	StephanieMoraAndrade	0.41514	5	Tue, 21 Feb 2017 18:54:45 (-0.4h)

Figure 6: Sexta puntuación obtenida en Kaggel