Integración, Limpieza, Validación y Análisis de Datos.

Tipología y Ciclo de Vida de los Datos: Práctica #2

Antonio Carcia

diciembre 31, 2018

Contents

Sumario

Fuente: www.kaggle.com
Insumo: *Consumo Gasolina Autos Ene 2018.xlsx*
Total Obs.: *4.617 tuplas*

Convenciones, formatos y nomenclatura:

La variable ***wd*** debe inicializarse con la ruta en la que se ubica el Fichero de Insumo.

Para propósitos de la estructura de directorios con la que suministra el ejercicio, se utiliza la ruta $HOME/dat/in.

El archivo con datos ajustados (procesados, listo para el análisis) se exporta a la ruta $HOME/dat/clean.

1. Introducción

Durante el siguiente trabajo se aborda un caso práctico en el contexto de la asignatura Tipología y Ciclo de Vida de Datos, dirigido a aplicar métodos y técnicas para la preparación de los datos, previamente a su uso en el desarrollo de proyectos analíticos.

Así mismo, se aplican criterios de análisis y selección de atributos, para efectos de la confección del set de datos final a ser utilizado en la construcción de modelos, durante el proceso de desarrollo.

Objetivos:

- Aplicar los conocimientos adquiridos para la resolución de problemas en entornos nuevos o poco conocidos.
- Identificar los datos relevantes y tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en ellos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.

2. Descripción del Dataset.

El juego de datos seleccionado para el trabajo está relacionado con el registro de características técnicas y prestaciones de marcas de vehículos con modelos hasta el 2.018, y su calificación en relación a la ***Emisión de Gases de Efecto Invernadero*** y potencial de ***Contaminación del Aire***. Este es de carácter público, y ha sido obtenido de la siguiente fuente:

URL: https://www.kaggle.com/checoalejandro/autos-consumo-gasolina-mexico#Consumo Gasolina Autos Ene 2018.xlsx

En general, la información relacionada con el set se presenta en forma resumida como sigue:

Aspecto	**Descripción**
Marca	Fabricante del vehículo
Submarca	Modelo de comercialización
Version	Versión de fabricación - breve descripción de características
Modelo	Año al que corresponde el modelo
Trans	Modelo de Transmisión específico o Tipo utilizado en la fabricación del vehículo
Comb	Tipo de Carburante o energía utilizada por el Motor del Vehículo
Cilindros	Cantidad de Cilindros que configuran el Motor
Potencia (HP)	Potencia del Motor
Tamaño (L)	Capacidad del Motor

Aspecto	Descripción
Categoría	Estilo o Clase de coche
R.Ciudad (Km/L)	Recorrido en Ciudad medido en km por litro de combustible
R.Carr (Km/L)	Recorrido en Carretera medido en km por litro de combustible
R.Comb (Km/L)	Recorrido en Km por Litro
R.Ajust (KM/L)	Recorrido Ajustado medido en Km por Litro
CO2 (gr/Km)	Emisiones de dióxido de carbono en gramos por Km recorrido
NOx (gr/1000Km)	Emisiones de óxido nitroso en gramos por unidades de 1000 Km
Calificación Gas Ef. Inv.	Calificación obtenida en medición de Gases de Efecto Invernadero emitidos
Calificación Contam. Aire	Calificación obtenida en relación al grado de Contaminación del Aire durante su desempeño

La muestra está formada por 4.617 observaciones, de 18 atributos.

Propósito:

Aunque el registro de marcas, versiones y modelos en bastante amplio, éste continuará creciendo en el tiempo, y con ello la necesidad de continuar registrando el impacto ambiental de cada uno de los nuevos modelos.

Se han observado imprecisiones en el juego de datos que afectan su integridad y completitud. Entre ellos, se aprecia una brecha importante en la columna de ***Calificación de Contaminación del Aire***.

Se propone desarrollar un modelo a partir de la caracterización recogida en este set, que permita calificar vehículos utilizando criterios similares a los implícitos en la muestra, completando en lo posible la brecha identificada en ellos.

Para esto será necesario procesar adecuadamente los datos objetos de estudio, identificar los atributos que más aportan en la descripción de la situación de interés, y que efectivamente pueden contribuir en la confección de un modelo a partir del que se facilite la correcta clasificación de los vehículos actualmente no calificados y nuevos por incorporar.

Si bien, la muestra disponible evidencia una condición de desbalanceo en la clasificación inicial que aportan, se espera lograr la confección de un modelo que resulte suficientemente explicativo y con capacidad predictiva, para contribuir en el objetivo que nos hemos impuesto.

De resultar adecuado el modelo que se obtenga, puede ser utilizado para completar los valores perdidos en relación a la calificación del renglón de ***Contaminación del Aire***, y soportar la validación de nuevos registros (producto de incorporaciones futuras).

3. Integración y Selección de los Datos de Interés.

El set de datos se asume autocontenido, por lo que no es necesario enriquecerlo, ni procesamientos adicionales en este sentido.

Se procede a la importación del archivo de insumo.

```
##### Cargando datos para exploración  --------------------
wd <- "./../dat/in"

setwd(wd)
cga2018 <- read.xlsx("Consumo Gasolina Autos Ene 2018.xlsx", sheet=1)

##### Total de filas cargadas  -----------------------------
nrow( cga2018 )
```

```
## [1] 4617
```

Una vez ejecutada la carga de datos, y luego de haber validado que la cantidad de observaciones incorporadas a las estructuras dispuestas por el ambiente de trabajo coincide con la documentación del set, se procede con la inspección de los tipos de datos asignados automáticamente a cada atributo durante la operación.

```
### Inspección de estructura de datos y verificación de tipos de datos asignados
###
str(cga2018)
```

```
## 'data.frame':    4617 obs. of  18 variables:
##  $ Marca                    : chr  "FORD" "FORD" "FORD" "FORD" ...
##  $ Submarca                 : chr  "FUSION" "FUSION" "FUSION" "FUSION" ...
##  $ Versión                  : chr  "HIBRIDO 4PTS 2.0L 4CIL 188HP AUT (CVT)" "HIBRIDO 4PTS 2.0L 4CIL
##  $ Modelo                   : chr  "2015" "2016" "2017" "2018" ...
##  $ Trans.                   : chr  "CVT" "CVT" "CVT" "AUT" ...
##  $ Comb.                    : chr  "Gasolina" "Gasolina" "Gasolina" "Gasolina" ...
##  $ Cilindros                : chr  "4" "4" "4" "4" ...
##  $ Potencia.(HP)            : num  188 188 188 188 156 156 90 90 90 110 ...
##  $ Tamaño.(L)               : num  2 2 2 2 2.5 2.5 1.5 1.5 1.5 1.3 ...
##  $ Categoría                : chr  "AUTOS COMPACTOS" "AUTOS COMPACTOS" "AUTOS COMPACTOS" "AUTOS COMPA
##  $ R..Ciudad.(km/l)         : num  27.4 27.4 25.6 25.6 24 ...
##  $ R..Carr..(km/l)          : num  28.6 28.6 24.8 24.8 21.9 ...
##  $ R..Comb..(km/l)          : num  28.9 28.9 25.2 25.2 23 ...
##  $ R..Ajust..(km/l)         : num  21.7 21.7 18.9 18.9 17.3 ...
##  $ CO2(g/km)                : num  107 107 123 123 135 135 119 119 127 134 ...
##  $ NOx.(g/1000km)           : num  5 0 2 2 5 6 7 7 7 10 ...
##  $ Calificación.Gas.Ef..Inv.: chr  "10" "10" "10" "10" ...
##  $ Calificación.Contam..Aire: chr  "9" "9" "9" "9" ...
```

Se observa que varios de los atributos responden al propósito de identificar la instancia a la que se refieren, exponiendo información de carácter descriptivo.

Tales campos serán convertidos a tipo factor para facilitar el procesamiento, tan pronto haya sido verificada la condición en relación a la ausencia de valores vacios. Particularmente, los campos ***Calificación Gases de Efecto Invernadero*** y ***Contaminación de Aire***, será convertidos a tipo entero, hasta que llegue la ocasión de formalizar nuevamente su conversión a tipo factor.

Así mismo, se aprovecha el ajuste para abreviar el nombre de algunos de los atributos del set

```
### Ajustes de tipos de atributos y denominación de columnas de la estructura
###
colnames(cga2018) <- c("marca",        # Fabricante del vehículo
                       "submarca",     # Submarca Comercial del vehículo
                       "vr",           # Versión del Modelo
                       "mod",          # Año al que corresponde el Modelo
                       "trans",        # Tipo de Transmisión
                       "comb",         # Tipo de Combustible que emplea
```

```
                        "cil",          # Número de Cilindros del motor
                        "pot",          # Potencia del motor
                        "tam",          # Tamaño o Capacidad ( en Ltros) del motor
                        "cat",          # Categoria del Vehículo
                        "r.Ciudad",     # Recorrido en Ciudad (Rendimiento por Litro)
                        "r.Carretera",  # Recorrido en Carretera (Rendimiento por Litro)
                        "r.Comb",       # Recorrido por Con¿mbustible (Rendimiento por Litro)
                        "r.Ajust",      # Recorrido Ajustado (Rendimiento por Litro)
                        "CO2",          # Emisión  de CO2 (g/km)
                        "NOx",          # Emisión  de NOX (g/1000km)
                        "GEI",          # Calificación: Emisión de Gases de Efecto Invernadero
                        "CA")           # Calificación: Contaminación del Aire

cga2018$GEI <- as.integer( cga2018$GEI )
```

Finalmente, el atributo ***Calificación de Contaminación del Aire***, también se considera un tipo enumerado, por lo que debe ser convertido a una variable categórica de tipo factor. Antes, una primera conversión a tipo entero, que sustitye los placeholders que indican valores perdidos sobre la variable (caracter '? ´) por valores N.A.

```
### Adecuación del Tipo correspondiente al atributo CA
###
cga2018$CA <- as.integer( cga2018$CA ) # Se convierten a NA los valores perdidos
```

```
## Warning: NAs introducidos por coerción
```

```
cga2018$CA <-factor( cga2018$CA )
```

4. Limpieza de Datos.

4.1. Elementos Vacíos, Ceros y Nulos

La validación de Valores tipo carácter con entradas vacías, demuestra que todas las entradas de este tipo se encuentran inicializadas, por lo que no se requiere tomar ninguna acción particular con respecto a este tipo de campos

```
### Verificación de Ausencia de Contenido
###
col.nbr <- colnames(cga2018)

for( i in 1:length(col.nbr) )
  if (is.character( cga2018[i] ))
    if ( sum(is.na(cga2018[i]) ) == 0)
       if ( sum( cga2018[i]== "" ) !=0 )
            print( col.nbr[i] )
```

En consecuencia, puede procederse con la conversión de tipos propuesta

```
### Conversión de Tipo para campos carácter
###
cga2018$mod   <-factor( cga2018$mod )
cga2018$trans <-factor( cga2018$trans )
cga2018$comb  <-factor( cga2018$comb )
cga2018$cat   <-factor( cga2018$cat )
```

```
cga2018$cil <- factor( cga2018$cil )
cga2018$cil <- ordered( cga2018$cil, levels=c("3","4","5", "6", "8", "10", "12"))
```

Continuando con la identificación de atributos con valores perdidos, sigue la verificación de presencia de valores N.A.

```
### Verificación de Valores Nulos
###
for(i in 1:length(col.nbr) )
   if ( sum(is.na(cga2018[i]) ) != 0){
        print( col.nbr[i] )
   }
```

```
## [1] "trans"
## [1] "CA"
```

Determinación de la cantidad de registros en situación de valores perdidos para el atributo ***trans***

```
### identificación de registros con Valores Nulos para atributo trans
###
which( is.na(cga2018$trans) )
```

```
##  [1]  182  183  363  364  365  397  454  455 1387 1388 3585 3586 3587 3588
## [15] 3589 3816
```

Consulta de contexto sobre registros con valores perdidos para el atributo ***trans***, con objeto de determinar algún criterio que permita decidir el tratamiento más adecuado para su inicialización:

```
### Consulta de registros con Valores Nulos para atributo trans
###
atr2show <- c("marca", "submarca", "vr", "mod")
cga2018[ which( is.na(cga2018$trans) ), atr2show ]
```

```
##      marca submarca                                        vr  mod
## 182  HONDA      FIT                FUN 5PTAS 1.5L 4CIL 130HP CVT 2017
## 183  HONDA      FIT                HIT 5PTAS 1.5L 4CIL 130HP CVT 2017
## 363  HONDA    CIVIC    EX COUPE 2PTAS 1.5L 4CIL 174HP TURBO CVT 2017
## 364  HONDA    CIVIC              SEDAN 4PTAS 1.5L 4CIL 174HP CVT 2017
## 365  HONDA    CIVIC              SEDAN 4PTAS 1.5L 4CIL 174HP CVT 2017
## 397  HONDA    CIVIC           EX SEDAN 4PTAS 2.0L 4CIL 158HP CVT 2017
## 454  HONDA     HR-V               EPIC 5PTAS 1.8L 4CIL 141HP CVT 2017
## 455  HONDA     HR-V               UNIQ 5PTAS 1.8L 4CIL 141HP CVT 2017
## 1387 HONDA   ACCORD           LX SEDAN 4PTAS 2.4L 4CIL 185HP CVT 2017
## 1388 HONDA   ACCORD              SPORT 4PTAS 2.4L 4CIL 189HP CVT 2017
## 3585 HONDA     CR-V            EXL 2WD 5PTAS 2.4L 4CIL 188HP CVT 2017
## 3586 HONDA     CR-V            EXL 4WD 5PTAS 2.4L 4CIL 188HP CVT 2017
## 3587 HONDA     CR-V             LX 2WD 5PTAS 2.4L 4CIL 188HP CVT 2017
## 3588 HONDA     CR-V TOURING 2WD 5PTAS 2.4L 4CIL 188HP TURBO CVT 2017
## 3589 HONDA     CR-V                2WD 5PTAS 2.4L 4CIL 188HP CVT 2017
## 3816 HONDA     CR-V             EX 2WD 5PTAS 2.4L 4CIL 184HP CVT 2017
```

Todos los vehículos a los que les falta valor para el campo ***trans*** son de ***marca*** **Honda** y ***modelo*** 2017, y de acuerdo al campo ***vr*** puede identificarse en otros registros similares que el valor faltante corresponde al tipo de caja **CVT**. Por lo tanto, el ajuste debe ser realizado asignando directamente dicho valor al atributo para aquellas observaciones que no exhiban valor (Esto es, se trata de una inconsistencia en los datos que puede ser corregida manualmente).

```
### Ajuste de registros con Valores Nulos para atributo trans
###
cga2018$trans <- as.character(cga2018$trans)
cga2018[ which( is.na(cga2018$trans) ), "trans" ] <- "CVT"
cga2018$trans <-factor( cga2018$trans )
```

La corrección de la lista de niveles incluidos en el factor ***trans***, requiere convertir el atributo a tipo carácter, realizar el ajuste, y finalmente reconvertir el atributo a tipo factor.

Siguiendo con el atributo destinado a la calificación por ***Contaminación de Aire*** (***CA***), tenemos que:

```
### Cuantificación de registros con Valores Nulos para atributo Calificación Contaminación Aire
###
length( which( is.na(cga2018$CA) ))
```

```
## [1] 1238
```

Este atributo exhibe una gran cantidad de valores perdidos (un poco más del 25% de las observaciones). Los elementos faltantes pueden intentar ser completados una vez construido el modelo, si las condiciones lo permiten.

No obstante, de las observaciones completas se dispone de valores para el atributo ***CA***, que serán tomados para efectos de confeccionar una partición que pueda ser empleada como set de Entrenamiento y otra como set de Test.

Identificación de atributos con valores inicializados a cero, que pudieran ser objetos de revisión.

```
### Verificación de atributos incorrectamente inicializaciones a Cero
###
for( i in 1:length(col.nbr) )
    if ( sum(is.na(cga2018[i]) ) == 0)
      if ( length(which(cga2018[i]== 0)) !=0 )
        print( col.nbr[i] )
```

```
## [1] "NOx"
## [1] "GEI"
```

En relación al atributo ***NOx***, en el siguiente segmento se aprecia que son muy pocos registros los que muestran afectación, existiendo observaciones de modelos/versiones previos del mismo fabricante, que exhiben valor para el atributo.

```
### Identificación y Consulta de registros con Valores incorrectamente inicalizados a Cero para atribut
###
cga2018[  which(cga2018$NOx == 0), atr2show]
```

```
##      marca submarca                                      vr  mod
## 2     FORD   FUSION    HIBRIDO 4PTS 2.0L 4CIL 188HP AUT (eCVT) 2016
## 2463  MINI COOPER S PACEMAN JCW 3PTS 1.6L 4CIL 218HP MAN TURBO 2016
## 2470  MINI COOPER S            3PTAS 1.6L 4CIL 218HP TURBO MAN 2017
```

```
head(cga2018[  which(cga2018$NOx != 0 & cga2018$submarca == "FUSION"), atr2show], 10)
```

```
##     marca submarca                                        vr  mod
## 1    FORD   FUSION        HIBRIDO 4PTS 2.0L 4CIL 188HP AUT (CVT) 2015
## 3    FORD   FUSION HIBRIDO 4X2 4PTAS 2.0L 4CIL 141(+47e)HP E-CVT 2017
## 4    FORD   FUSION        HIBRIDO 4PTAS 2.0L 4CIL 188HP AUT eCVT 2018
## 5    FORD   FUSION               HYBRID 4PTS 2.5L 4CIL 156HP CVT 2011
## 6    FORD   FUSION               HYBRID 4PTS 2.5L 4CIL 156HP CVT 2012
## 414  FORD   FUSION                      4PTS 2.5L 4CIL 170HP AUT 2014
```

```
## 415  FORD   FUSION               4PTS 2.5L 4CIL 168HP AUT (6F35) 2015
## 416  FORD   FUSION               4PTS 2.5L 4CIL 175HP AUT (6F35) 2016
## 430  FORD   FUSION                 4X2 4PTAS 2.5L 4CIL 175HP AUT 2017
## 431  FORD   FUSION                     4PTAS 2.5L 4CIL 175HP AUT 2018
```

```
head(cga2018[  which(cga2018$NOx != 0 & cga2018$submarca == "COOPER S"), atr2show], 10)
```

```
##      marca submarca                                        vr  mod
## 2286  MINI COOPER S       HATCH 5PTAS 2.0L 4CIL 192HP AUT TURBO 2016
## 2287  MINI COOPER S   HATCHBACK 5PTAS 2.0L 4CIL 192HP TURBO AUT 2017
## 2292  MINI COOPER S CONVERTIBLE 2PTAS 2.0L 4CIL 192HP AUT TURBO 2016
## 2293  MINI COOPER S             5PTAS 2.0L 4CIL 189HP TURBO AUT 2017
## 2301  MINI COOPER S              3PTS 2.0L 4CIL 192HP AUT TURBO 2016
## 2302  MINI COOPER S             3PTAS 2.0L 4CIL 192HP TURBO AUT 2017
## 2303  MINI COOPER S  COUNTRYMAN 5PTAS 1.6L 4CIL 190HP MAN TURBO 2016
## 2304  MINI COOPER S     CLUBMAN 6PTAS 2.0L 4CIL 192HP AUT TURBO 2016
## 2306  MINI COOPER S CONVERTIBLE 2PTAS 2.0L 4CIL 192HP TURBO AUT 2017
## 2307  MINI COOPER S       COUPE 2PTAS 2.0L 4CIL 231HP TURBO AUT 2017
```

Lo anterior nos induce a pensar que el caso se refiere a un valor perdido (inicializado con el centinela de valor cero), y que debe ser corregido. Para ello se utilizará más adelante una aproximación tipo **kNN**.

Haciendo extensivo el analisis al atributo ***GEI***, aplicando el mismo reazonamiento ilustrado anteriormente para el atributo ***NOx***, resulta sencillo cuantificar el número de registros afectados por la presencia de valores perdidos para dicho campo.

```
### Cuantificación de registros con Valores incorrectamente inicializaciones a Cero (GEI)
###
length( which( cga2018$GEI == 0 & cga2018$NOx != 0  & cga2018$CO2 != 0 ) )
```

```
## [1] 411
```

La existencia de observaciones con modelos y versiones previos del mismo fabricante (con valores distintos de cero para el atributo), aunado a que se tienen observaciones con el atributo inicializado a 0.00, que exhiben valor para el atributo ***CA***, inducen a pensar de que este tipo de casos constituyen una contradicción (inconsistencia).

En el segmento que sigue se adelanta una muestra de la consulta que apoya la argumentación, así como de la cuantificación de los registros afectados por la inicialización a cero para el atributo ***GEI***, en ocasión de ausencia y presencia de valor en el atributo ***CA***, respectivamente.

```
### Consulta de registros con Valores incorrectamente inicializados a Cero para atributo cga2018$GEI
###
atr2show <- c(atr2show,  "CA")
head( cga2018[  which(  cga2018$GEI == 0
                      & cga2018$NOx != 0
                      & cga2018$CO2 != 0
                      & cga2018$CA != 0 ), atr2show],
      10)
```

```
##        marca submarca                                          vr  mod CA
## 1912 BENTLEY MULSANNE 4PTS 6.8L 8CIL 505HP TIP DOBLE TURBO QUATTRO 2013  9
## 1913 BENTLEY MULSANNE  4PTS 6.8L 8CIL 505HP TIP QUATTRO DOBLETURBO 2014  9
## 2001   DODGE      300               SRT-8 4PTS 6.4L 8CIL 465HP AUT 2012  7
## 2002   DODGE  CHARGER               SRT-8 4PTS 6.4L 8CIL 470HP AUT 2012  7
## 2003   DODGE      300               SRT-8 4PTS 6.4L 8CIL 465HP AUT 2013  7
## 2004   DODGE  CHARGER               SRT-8 4PTS 6.4L 8CIL 470HP AUT 2013  7
## 2013    AUDI       R8            2PTS 5.2L 10CIL 525HP MAN QUATTRO 2016  8
```

```
## 2014    AUDI       R8        2PTS 5.2L 10CIL 525HP STRONIC QUATTRO 2016  8
## 2015    AUDI       R8        2PTS 5.2L 10CIL 550HP STRONIC QUATTRO 2016  8
## 2017   DODGE  CHARGER                 SRT4PTAS 6.2L 8CIL 707HP AUT 2016  7
```

```
### Cuantificación de registros con Valores inicializados a Cero para atributo cga2018$GEI
### en presencia de valores asignados al atributo cga2018$CA
###
length( which( cga2018$GEI == 0 & (cga2018$NOx != 0 | cga2018$CO2 != 0)
                                & (is.na(cga2018$CA) != TRUE ) ) )
```

```
## [1] 228
```

```
### Cuantificación de registros con Valores inicializados a Cero para atributo cga2018$GEI
### en ausencia de valor en el atributo cga2018$CA
###
length( which( cga2018$GEI == 0 & (cga2018$NOx != 0 | cga2018$CO2 != 0)
                                & (is.na(cga2018$CA) == TRUE ) ) )
```

```
## [1] 183
```

El número de observaciones con el atributo ***GEI*** igual a cero es de 411 registros, que representa algo menos el 10% de la muestra. Aunque es una proporción significativa desde el punto de vista del tamaño de nuestro juego de datos, tiene sentido tratar de imputar el valor.

Los datos del ejercicio están clasificados por una serie de atributos categóricos, que junto con consideraciones propias del dominio del problema, pueden ser de valor a los efectos de aplicar algún procedimiento para la imputación de valores.

En este sentido, parece razonable tratar de resolver estos valores para el atributo, teniendo en cuenta información del Fabricante (***marca***) y del producto (***submarca***), combinados con otras variables incluidas en la muestra.

Siguiendo el lineamiento adelantado en relación a la resolución de valores perdidos para el atributo ***NOx***, se utilizará igualmente la aproximación **kNN** para la resolución requerida por el atributo ***GEI***.

El siguiente segmento aplica el procedimiento, creando previamente una copia de los datos y suprimiendo las columnas que podrían introducir ruido en la resolución, a causa del nivel de especificidad que introducen en el conjunto de datos.

```
### Resolución de valores perdidos (inicializados a Cero) - atr NOx y GEI
###

# Supresión de Versión, Modelo y Calificación Contaminación Aire
cga2018.dat <- cga2018[ c(-3,-4,-18)]

# Adecuación de datos para el procedimiento : Sustitución de Ceros por NA
cga2018.dat[  which(cga2018$NOx == 0), "NOx"] <- NA

cga2018.dat[  which(  cga2018$GEI == 0
                    & cga2018$NOx != 0
                    & cga2018$CO2 != 0 ), "GEI"] <- NA

# Ejecución del procedimiento con un número impar de Vecinos (k)
cga2018.imp <- kNN(cga2018.dat, k=5)
```

Verificación de atributos con ajustes sugeridos por el método de imputación

```
### Verificación de columnas con valores de imputación calculadas con kNN
###
```

```
col.nbr <- col.nbr[ c(-3,-4,-18)] # Supresión de Versión, Modelo y Calificación Contaminación Aire

### La matriz de resulados de la imputación dobla la cantidad de columnas originales
### por lo que se reajustan los índices para el recorrido adecuado de las columnas
##
desde <- length(col.nbr) + 1
hasta <- length(col.nbr) * 2

for(i in desde:hasta )
   if ( sum( cga2018.imp[i] == TRUE ) != 0 )
      print( col.nbr[i - length(col.nbr)] )
```

```
## [1] "NOx"
## [1] "GEI"
```

Traslación de los estimados obtenidos para los atributos ***NOx*** y ***GEI***, inicializados con centinelas de valor cero.

```
### Imputación de los atributos NOx y GEI
###
copiar.imp <- which(cga2018.imp$NOx_imp == TRUE)
cga2018$NOx[ copiar.imp ] <- cga2018.imp$NOx[ copiar.imp ]

copiar.imp <- which(cga2018.imp$GEI_imp == TRUE)
cga2018$GEI[ copiar.imp ] <- cga2018.imp$GEI[ copiar.imp ]
```

Habiendo concluido la imputación de la variable ***GEI***, resulta conveniente proceder a su conversión al tipo factor, tal como se adelantaba en las secciones iniciales.

```
### Conversión del atributo GEI a tipo factor
###
cga2018$GEI <- factor( cga2018$GEI )
```

4.2 Reducción por Simplificación de Valores

El factor representado por la variable ***trans*** puede ser objeto de una simplificación importante al sustituir los tipos específicos y modelos de transmisión, por la categoría a la que pertenece cada modelo incluido en el set de datos. De esta manera, reduciremos el número de niveles del factor de alrededor de 20 a tan solo 5.

Para ello nos apoyamos en una breve investigación, que nos permite resumir el conjunto de transmisiones utilizadas en el set de datos a sus tipos generales:

- Manual,
- Automática,
- Manual Robotizada,
- Variación Continua, y
- Doble Embrague

```
### Reducción del factor trans con los Tipos de Transmisión Generales
###

cga2018$trans <- as.character(cga2018$trans)

cga2018[ which( cga2018$trans == "ASG"
              | cga2018$trans == "DUALOGIC"
              | cga2018$trans == "G TRONIC"
```

```
              | cga2018$trans == "R TRONIC"
              | cga2018$trans == "S TRONIC"
              | cga2018$trans == "S TRONIC 7"
              | cga2018$trans == "STRONIC"
             ),
        "trans" ] <- "Manual.Robot"

cga2018[ which( cga2018$trans == "AUT"
              | cga2018$trans == "TIP"
              | cga2018$trans == "TIP 8"
              | cga2018$trans == "TIPT"
              | cga2018$trans == "TIPTRONIC"
              | cga2018$trans == "ZF SPEED QUICKSHIFT"
             ),
        "trans" ] <- "Auto"

cga2018[ which( cga2018$trans == "CVT"
              | cga2018$trans == "MULTIT"
              | cga2018$trans == "MULTITRONIC"
             ),
        "trans" ] <- "Var.Continua"

cga2018[ which( cga2018$trans == "M5"
              | cga2018$trans == "M6"
              | cga2018$trans == "MAN 6"
              | cga2018$trans == "MAN"
             ),
        "trans" ] <- "Manual"

cga2018[ which( cga2018$trans == "DCT"
              | cga2018$trans == "DKG"
              | cga2018$trans == "DSG"
              | cga2018$trans == "PDK"
            ),
        "trans" ] <- "Dbl.Embrague"

cga2018$trans <-factor( cga2018$trans )

## Niveles resultantes para el factor
levels(cga2018$trans)
```

```
## [1] "Auto"         "Dbl.Embrague" "Manual"       "Manual.Robot"
## [5] "Var.Continua"
```

4.3. Identificación y Tratamiento de Valores Extremos

A título de referencia se describe cuantitativamente la muestra.

```
### Explorando algunas estadisticas asociadas a los atributos de la muestra
###
summary(cga2018$pot)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      60     150     220     255     330     888
```

```r
summary(cga2018$tam)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.898   1.800   2.500   2.870   3.600   8.400
```

```r
summary(cga2018$r.Ciudad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.10    8.20   10.42   10.60   12.81   27.46
```

```r
summary(cga2018$r.Carretera)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.70   13.48   16.39   16.61   19.60   31.30
```

```r
summary(cga2018$r.Comb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.96   10.46   12.87   13.18   15.61   28.93
```

```r
summary(cga2018$r.Ajust)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.720   7.850   9.650   9.888  11.710  21.700
```

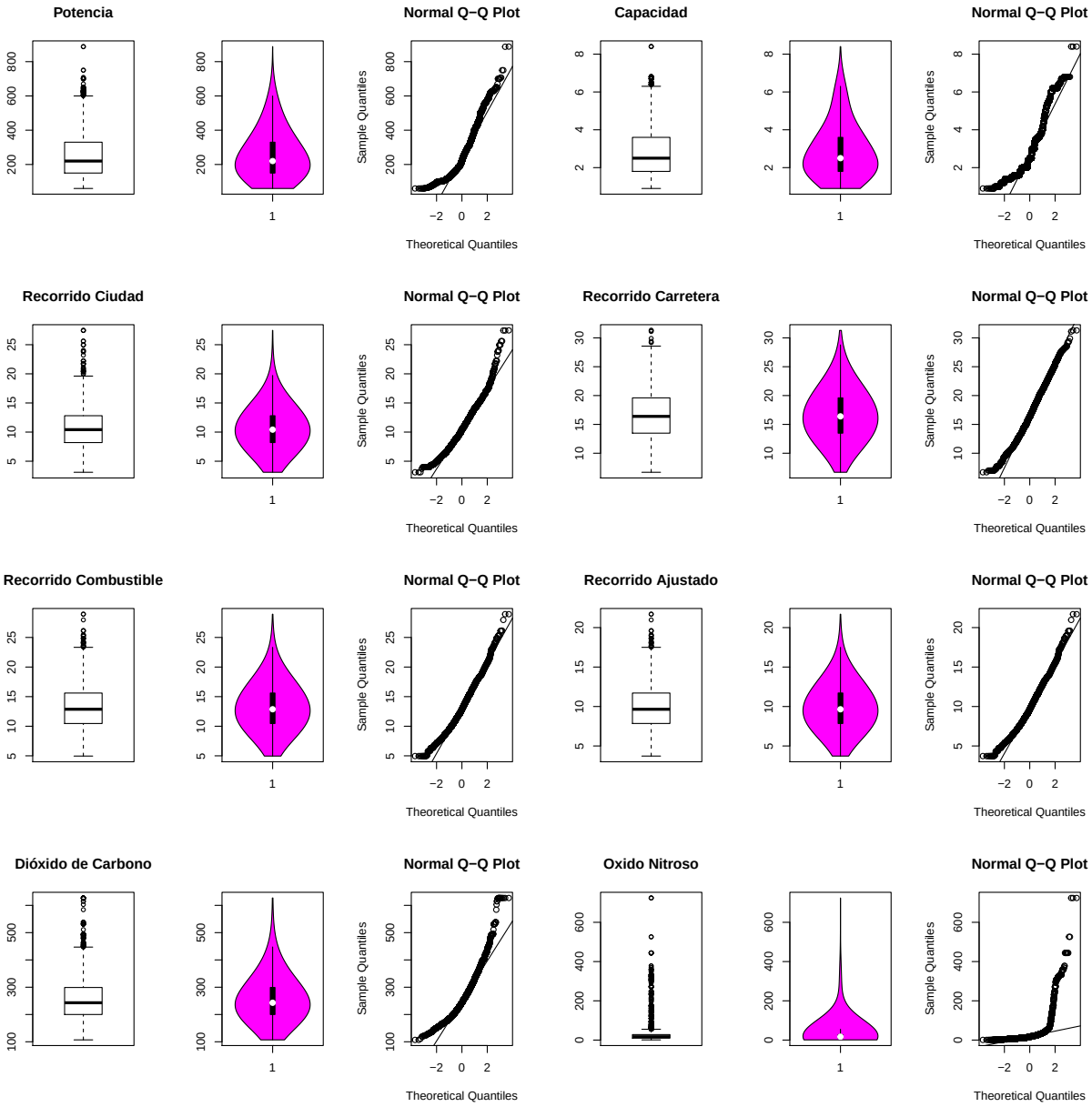
```r
summary(cga2018$CO2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   107.0   200.0   243.0   256.5   299.0   627.0
```

```r
summary(cga2018$NOx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   17.00   30.81   28.00  724.00
```

Siendo relativamente pocos los atributos numéricos que componen la muestra, podemos graficar para cada uno de ellos diagramas de caja, de violín y de cuantiles vs cuantiles. La idea es tratar de visualizar la presencia de outliers, la forma aproximada de las distribuciones en la que se presentan los valores de cada campo, y facilitar una identificación preliminar de la presencia de posibles distribuciones normales.



Los diagramas de violín que corresponde a cada variable nos permiten apreciar su forma aproximada, y darnos una idea acerca de su simetría En nuestro caso, todas sin excepción aparecen con alguna punta truncada, por lo que la normalidad de las distribuciones pudiera estar bastante cuestionada.

Esto se corrobora con los gráficos "quantile vs quantile" (qq) que se ha incorporado para cada variable. En ellos se observa que en el mejor de los casos, la adherencia a la diagonal se afecta por distorsiones en los extremos o en el centro.

Lo anterior nos obliga a validar la normalidad de estas variables con técnicas numéricas, aspecto que se aborda en sucesivos apartes.

En otro orden de ideas, las colas que se aprecian en los diagramas de caja, evidencian la presencia de valores extremos; siendo los casos con mayor ocurrencia los de los atributos ***CO2*** y ***NOx***, tal como se muestra de seguido.

```
##### Número de observaciones fuera del intervalo definido por 1.5*IQR para cada variable numérica
##
out.Vi <- c(  length(out.pot),    length(out.tam),     length(out.r.Ciudad), length(out.r.Carretera),
              length(out.r.Comb), length(out.r.Ajust), length(out.CO2),      length(out.NOx) )
out.Vi
```

```
## [1]  73  58  36   7  33  33  90 342
```

En el ejemplo que nos ocupa, se reportan diversos tipos de vehículos (modelos dirigidos al mercado masivo, así como otros de alta gama), obviamente con características distintas. Particularmente los modelos de lujo de marcas prominentes, ofrecen versiones con prestaciones muy por encima de las utilizadas por los modelos más económicos.

Esta segmentación es la primera causa de diferencias tan pronunciadas en los casos identificados, en atributos como la ***potencia*** y la ***capacidad*** del motor (***tamaño***), que a su vez repercuten sobre el rendimiento del vehículo en relación al consumo de combustible.

Así mismo, una breve revisión en relación a la emisión de ***CO2*** y ***NOx***, nos permite apreciar que los motores **Diesel** son los que en promedio reportan altos valores para estos atributos, seguidos luego por modelos gran ***potencia*** a ***Gasolina***.

Por lo anterior se puede argüir que por razones del domino del problema, podría ser contraproducente hacer conjeturas en cuanto a la conveniencia y validez de las acciones para resolver los casos de "outliers" bajo sospecha (por la inspección anterior).

Luego, una primera aproximación (conservadora) sugiere asumirlos como válidos; por lo tanto, una alternativa es la de no descartarlos o imputarlos, y proseguir con el ejercicio.

<<< Atención>>> Más adelante se reconsidera la necesidad o conveniencia de mantener el atributo ***Combustible***, cuya resolución incide favorablemente en la atenuación de la eventual distorsión producto de la presencia de outliers en los campos ***NOx*** y ***CO2***.

4.4 Exportación de Datos Procesados

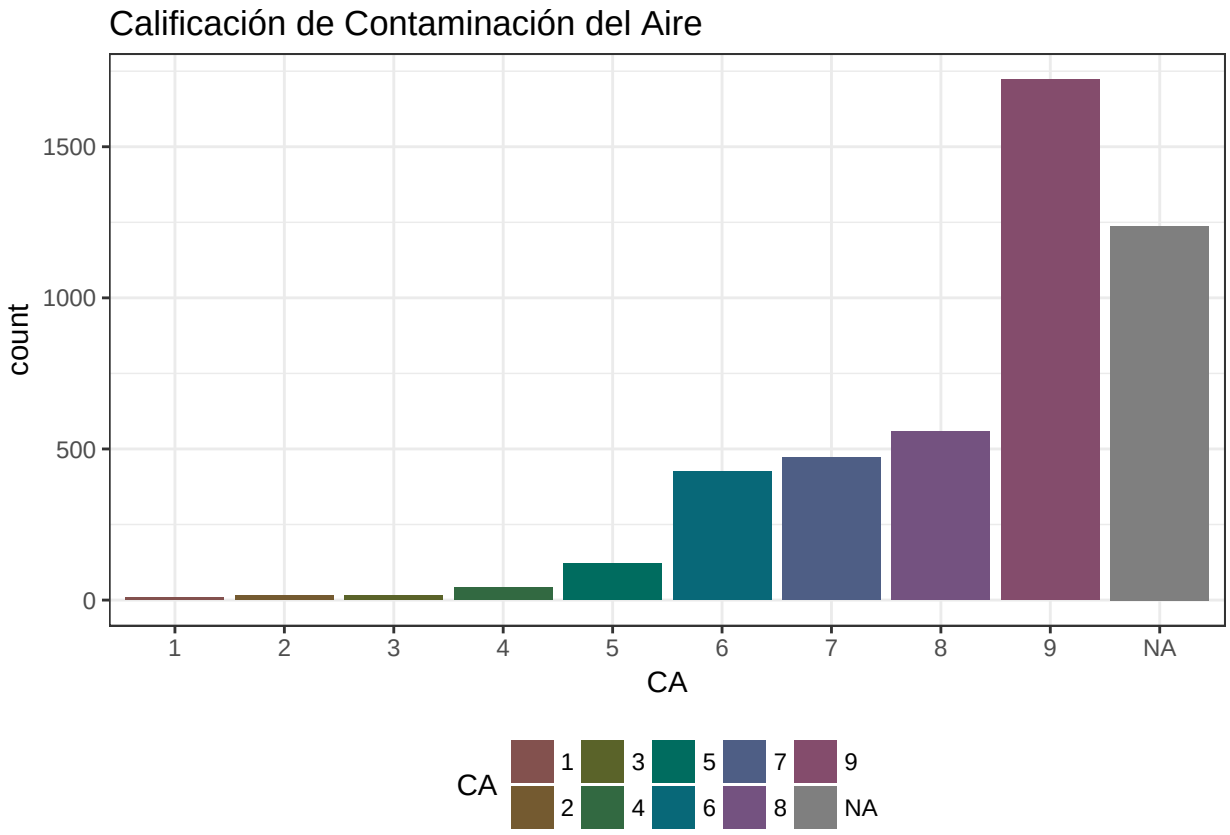
Realizadas las correcciones y ajustes que se han estimado adecuadas para la preparación de los datos, antes de proceder con su procesamiento, se sigue la exportación de una versión depurada.

```
#### Exportación de datos depurados
##
write.csv2(cga2018, "./../dat/clean/Consumo Gasolina Autos Ene 2018 - clean.csv")
```

4.5 Distribución de la Variable de Respuesta

Con el objetivo de confeccionar un modelo que permita la clasificación de observaciones, en categorías correspondientes a cada nivel del factor de calificación de ***Contaminación del Aire***, se dedica esta sección a revisar brevemente cómo se distribuye la variable ***CA*** dentro de la muestra, y con respecto a otros atributos.

```
### Distribución de la variable de Respuesta CA
###
ggplot(data = cga2018, aes(x = CA, y = ..count.., fill = CA)) +
  geom_bar() +
  scale_fill_hue(l=40, c=35) +
  labs(title = "Calificación de Contaminación del Aire") +
  theme_bw() +
  theme(legend.position = "bottom")
```



Acompañamos la gráfica con las frecuencias observadas y las proporciones en que ocurre cada nivel del factor

```
### Tabla de frecuencias de la variable de Respuesta CA
###
table(cga2018$CA)
```

```
##
##    1    2    3    4    5    6    7    8    9
##    8   16   14   41  122  426  471  559 1722
```

```
prop.table(table(cga2018$CA)) %>% round(digits = 4)
```

```
##
##      1      2      3      4      5      6      7      8      9
## 0.0024 0.0047 0.0041 0.0121 0.0361 0.1261 0.1394 0.1654 0.5096
```

El resultado revela un fuerte desequilibrio en relación a la forma en que se distribuyen las proporciones de la variable calificación de ***Contaminación de Aire*** sobre la muestra, concentrando la mayor cantidad de marcas, modelos y versiones en las categorías más altas.

4.5.1 Simplificación de la Variable de Respuesta

Para la variable de respuesta **CA** se tiene una distribución, en la que los valores iniciales de la escala (hasta el 5) prácticamente no se diferencian (siendo su volumen bastante bajo); así como también, para los niveles del segmento de la escala que van del valor 6 al 8, los valores observados son igualmente homogéneos.

Considerando esta distribución particular, parece razonable proponer la simplificación de la escala para la variable, consolidando los valores de manera de contar tan solo con tres niveles, de acuerdo al siguiente

arreglo:

- Grupo de Contaminación de Aire **Baja**, integrada por las observaciones con valores del 1 al 5.
- Grupo de Contaminación de Aire **Moderada**, que incluye las que exhiben valores entre 6 y 8.
- Grupo de Contaminación de Aire **Alta**, para las observaciones con valor 9.

El siguiente segmento, ejecuta la simplificación propuesta:

```
### Reducción del factor CA con la convención: Baja (1), Moderada (2) y Alta (3)
###

cga2018$CA <-as.integer( cga2018$CA )

cga2018[ which( cga2018$CA >= 1 & cga2018$CA <= 5 ), "CA" ] <- 1
cga2018[ which( cga2018$CA >= 6 & cga2018$CA <= 8 ), "CA" ] <- 2
cga2018[ which( cga2018$CA == 9                    ), "CA" ] <- 3

cga2018$CA <-factor( cga2018$CA )
```

Donde se representa con el valor 1 la clase de Contaminación de Aire **Baja**, con el valor 2 la **Moderada**, y finalmente la **Alta** con el valor 3.

4.6 Distribución de Variables Categóricas

El siguiente aparte está dirigido a visualizar cómo se distribuye la variable de respuesta con respecto a cada una de las variables categóricas de la muestra, que han sido consideradas como con potencial para influir sobre el resultado de un eventual modelo de clasificación.

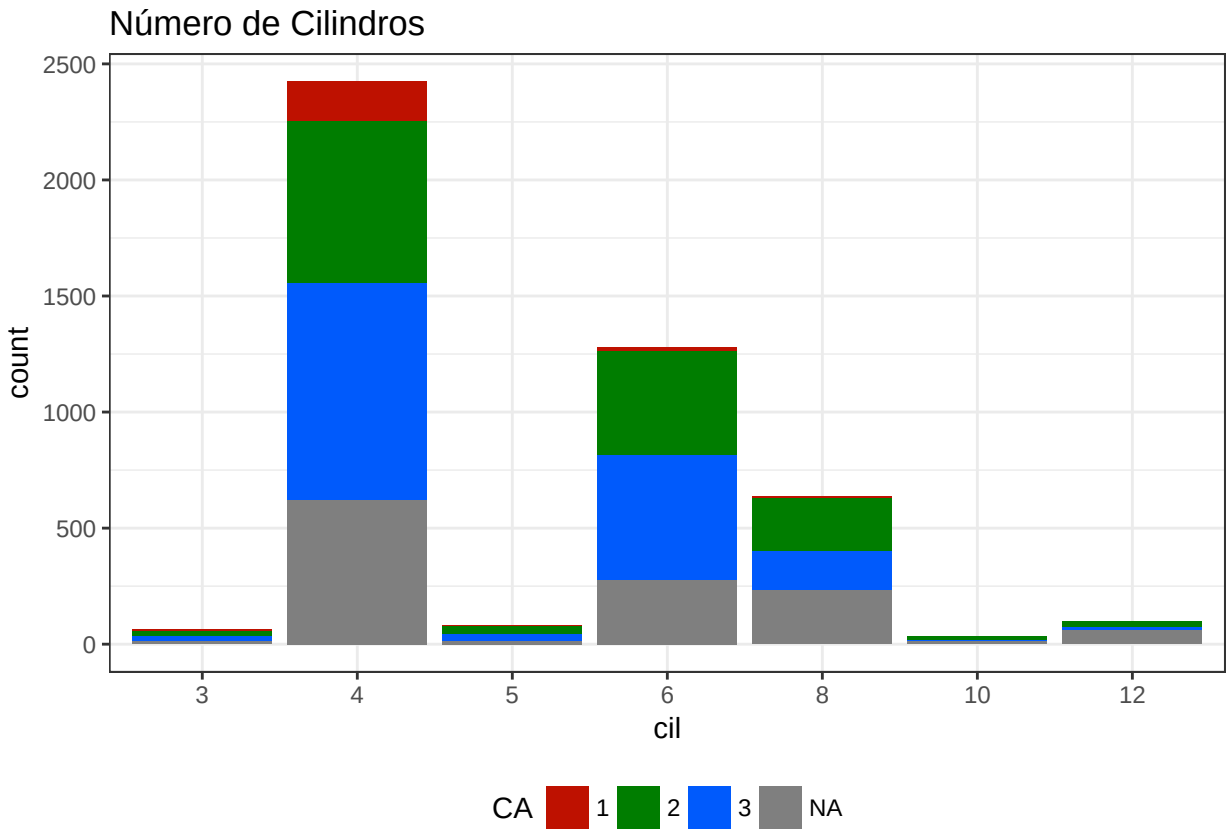
```
### Gráfica de la distribución del atributo CA sobre el Número de Cilindros
###

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
ggplot(data = cga2018, aes(x = cil, y = ..count.., fill = CA)) +
              geom_bar() +
              scale_fill_hue(l=40, c=135) +
              labs(title = "Número de Cilindros") +
              theme_bw() +
              theme(legend.position = "bottom")
```



El mayor volumen de modelos se concentra en la categoría de cuatro cilindros, seguidas a distancia por los segmentos de seis y ocho cilindros. En general no se aprecia una cobertura uniforme sobre las categorías de la variable ***Contaminación del Aire***.

Para facilitar la interpretación de la gráfica, se anexa la tabla de frecuencia y proporciones correspondientes a cada categoría:

```
### Frecuencia y Proporciones de la distribución del atributo CA sobre la variable cil
###
# Tabla de frecuencias relativas a la Contaminación del Aire por Cilindro
table(cga2018$cil,
      cga2018$CA,
      dnn=c("# Cilindros","Contaminación del Aire"))
```

```
##            Contaminación del Aire
## # Cilindros   1   2   3
##          3    6  19  22
##          4  170 694 938
##          5    1  33  32
##          6   17 446 541
##          8    7 225 169
##          10   0  16   3
##          12   0  23  17
```

```
# Tabulación de proporciones
prop.table(table(cga2018$cil, cga2018$CA), margin = 1) %>% round(digits = 2)
```

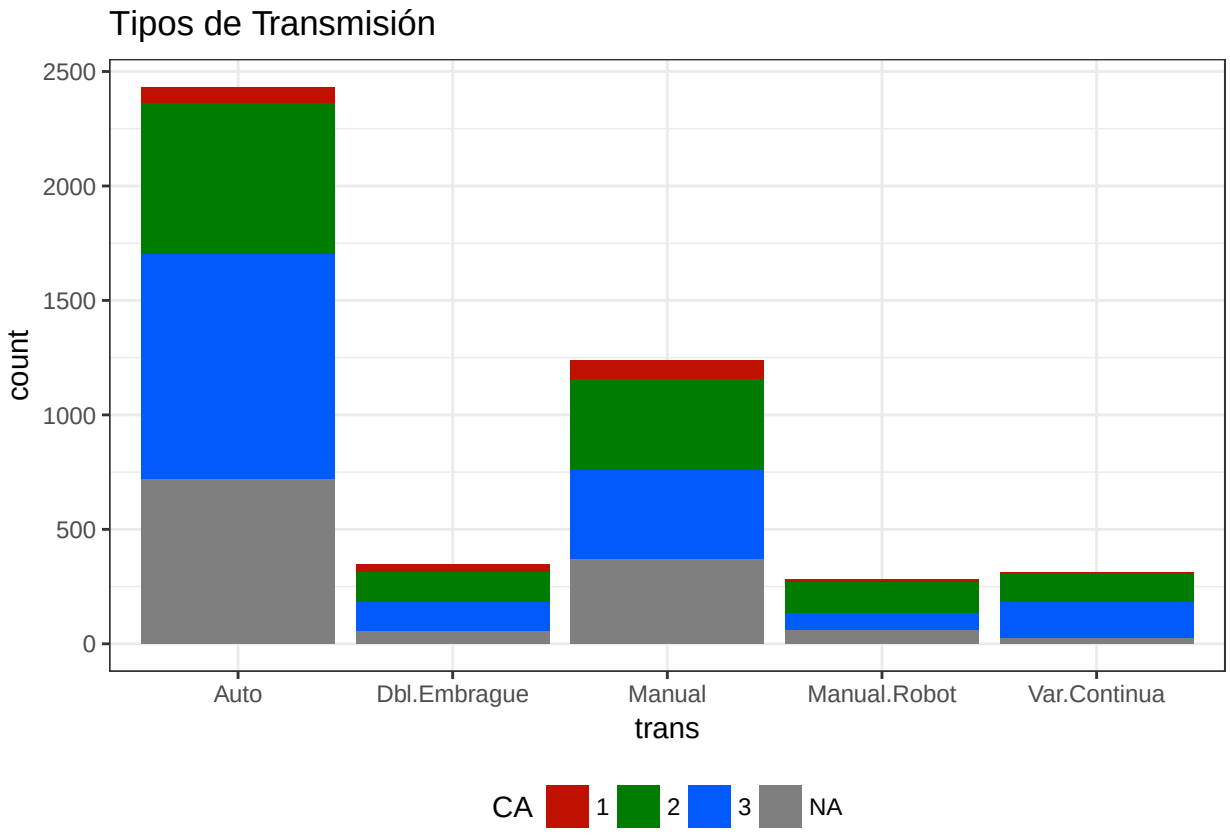
```
##
```

```
##         1    2    3
##   3  0.13 0.40 0.47
##   4  0.09 0.39 0.52
##   5  0.02 0.50 0.48
##   6  0.02 0.44 0.54
##   8  0.02 0.56 0.42
##   10 0.00 0.84 0.16
##   12 0.00 0.57 0.42
```

Para el caso de la variable relacionada al ***Número de Cilindros*** se hace evidente el desequilibrio de la distribución por categoría de ***Contaminación de Aire*** (claramente sesgado hacia los valores altos de la clasificación).

La gráfica por Tipo de Transmisión, presenta un comportamiento similar a la anterior, con una concentración mayor de modelos construidos con cajas de cambio Automáticas; pero con una distribución más proporcionada sobre las categorías de ***Contaminación del Aire***.

```
### Gráfica de la distribución del atributo CA sobre el Tipos de Transmisión
###
ggplot(data = cga2018, aes(x = trans, y = ..count.., fill = CA)) +
               geom_bar() +
               scale_fill_hue(l=40, c=135) +
               labs(title = "Tipos de Transmisión") +
               theme_bw() +
               theme(legend.position = "bottom")
```



Sigue la tabulación de frecuencias y proporciones observadas en la muestra:

```
### Frecuencia y Proporciones de la distribución del atributo CA sobre la variable trans
###
# Tabla de frecuencias relativas a la Contaminación del Aire por Tipo de Transmisión
table(cga2018$trans,
      cga2018$CA,
      dnn=c("Tipo de Transmisión","Contaminación del Aire"))
```

```
##                    Contaminación del Aire
## Tipo de Transmisión   1   2   3
##        Auto          69 662 981
##        Dbl.Embrague  31 133 124
##        Manual        80 398 388
##        Manual.Robot  13 139  72
##        Var.Continua   8 124 157
```

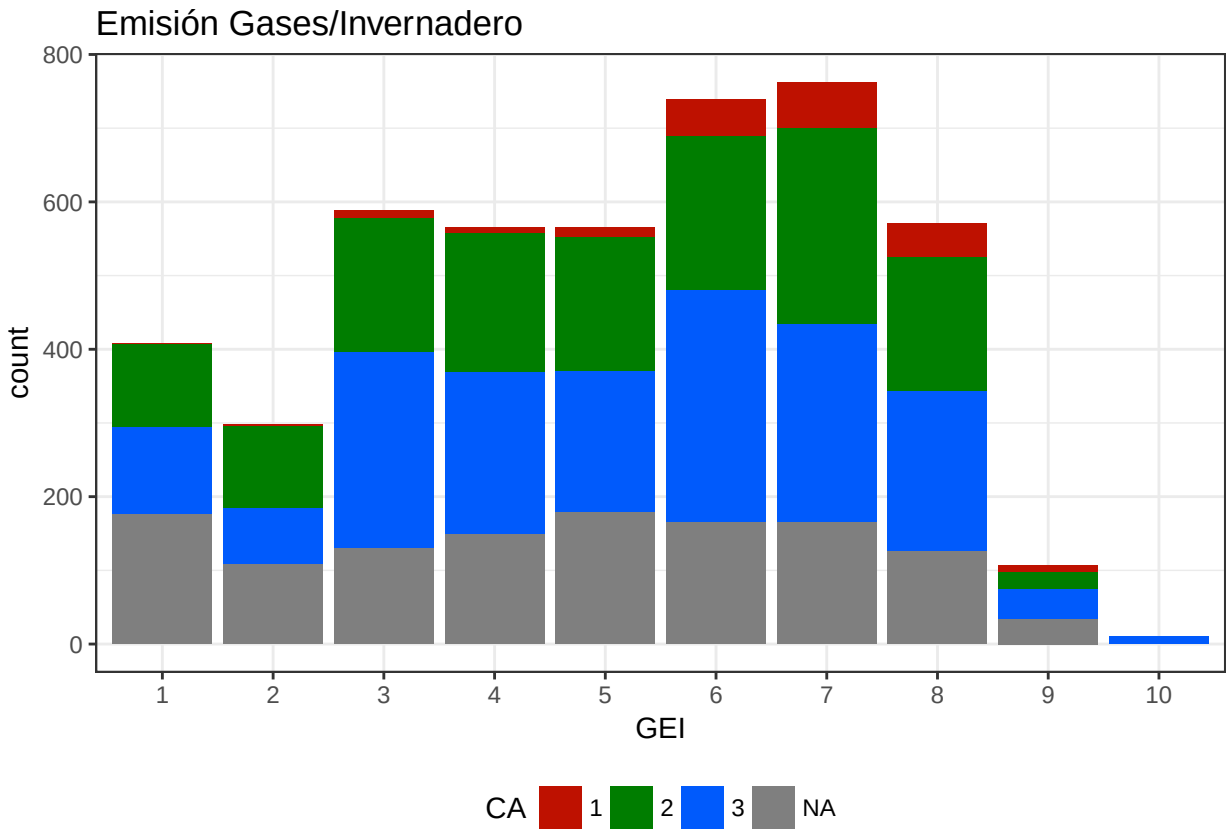
```
# Tabulación de proporciones
prop.table(table(cga2018$trans, cga2018$CA), margin = 1) %>% round(digits = 2)
```

```
##
##                    1    2    3
##   Auto          0.04 0.39 0.57
##   Dbl.Embrague 0.11 0.46 0.43
##   Manual       0.09 0.46 0.45
##   Manual.Robot 0.06 0.62 0.32
##   Var.Continua 0.03 0.43 0.54
```

En el caso del ***Tipo de Transmisión***, aunque persiste el desbalance en la distribución con respecto a las categorías de ***Contaminación del Aire***, el efecto se percibe más atenuado.

Para cerrar la sección se presenta la gráfica correspondiente a la variable de ***Calificación de Emisión de Gases de Efectos Invernadero***

```
### Gráfica de la distribución del atributo CA sobre la Calificación de Emisión de Gases EI
###
ggplot(data = cga2018, aes(x = GEI, y = ..count.., fill = CA)) +
              geom_bar() +
              scale_fill_hue(l=40, c=135) +
              labs(title = "Emisión Gases/Invernadero") +
              theme_bw() +
              theme(legend.position = "bottom")
```



Esta variable se aprecia mejor distribuida en el rango de categorías de la variable ***Contaminación del Aire***; sin embargo, debido a su carácter estructural, el desbalance en la distribución persiste como puede apreciarse en la poca presencia sobre la categoria ***CA* Baja** (1) para niveles bajos de **Emisión de GEI**, que se confirma en las tabulaciones de frecuencia y proporciones que sigue:

```
### Frecuencia y Proporciones de la distribución del atributo CA sobre la variable GEI
###
# Tabla de frecuencias relativas a la Contaminación del Aire por categoria de Emisión de GeI
table(cga2018$GEI,
      cga2018$CA,
      dnn=c("Emisión de Gases EI","Contaminación del Aire"))
```

```
##                     Contaminación del Aire
## Emisión de Gases EI   1   2   3
##                   1    1 112 119
##                   2    2 111  76
##                   3   10 181 266
##                   4    8 189 219
##                   5   13 183 190
##                   6   50 209 315
##                   7   62 267 268
##                   8   46 181 218
##                   9    9  23  41
##                   10   0   0  10
```

```
# Tabulación de proporciones
prop.table(table(cga2018$GEI, cga2018$CA), margin = 1) %>% round(digits = 2)
```

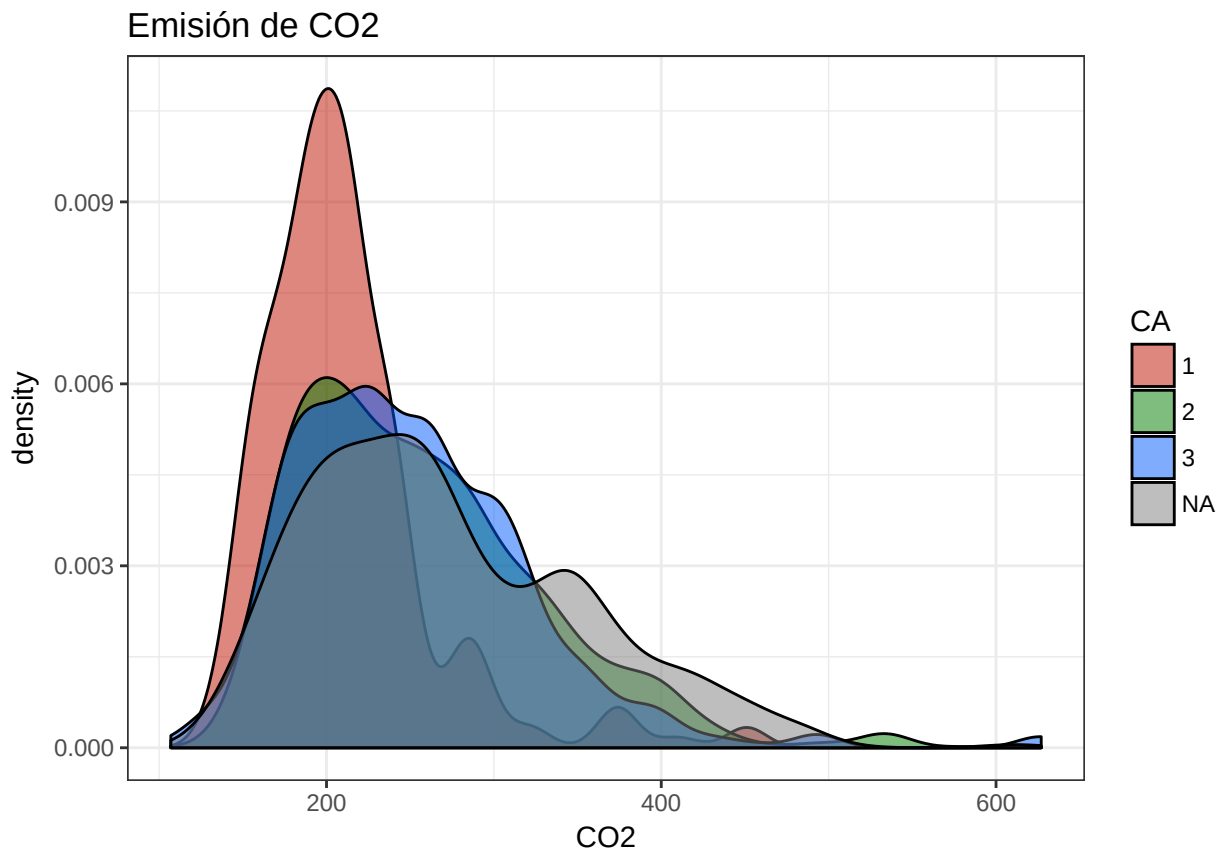
```
##
##         1    2    3
##   1  0.00 0.48 0.51
##   2  0.01 0.59 0.40
##   3  0.02 0.40 0.58
##   4  0.02 0.45 0.53
##   5  0.03 0.47 0.49
##   6  0.09 0.36 0.55
##   7  0.10 0.45 0.45
##   8  0.10 0.41 0.49
##   9  0.12 0.32 0.56
##   10 0.00 0.00 1.00
```

4.7 Distribución de Variables Continuas

Se presume la existencia de una relación (de mucha relevancia) entre la variable de respuesta para la calificación de la ***Contaminación del Aire***, y los atributos continuos relacionados a la medición de emisiones a la atmosfera (***CO2*** y ***NOx***).

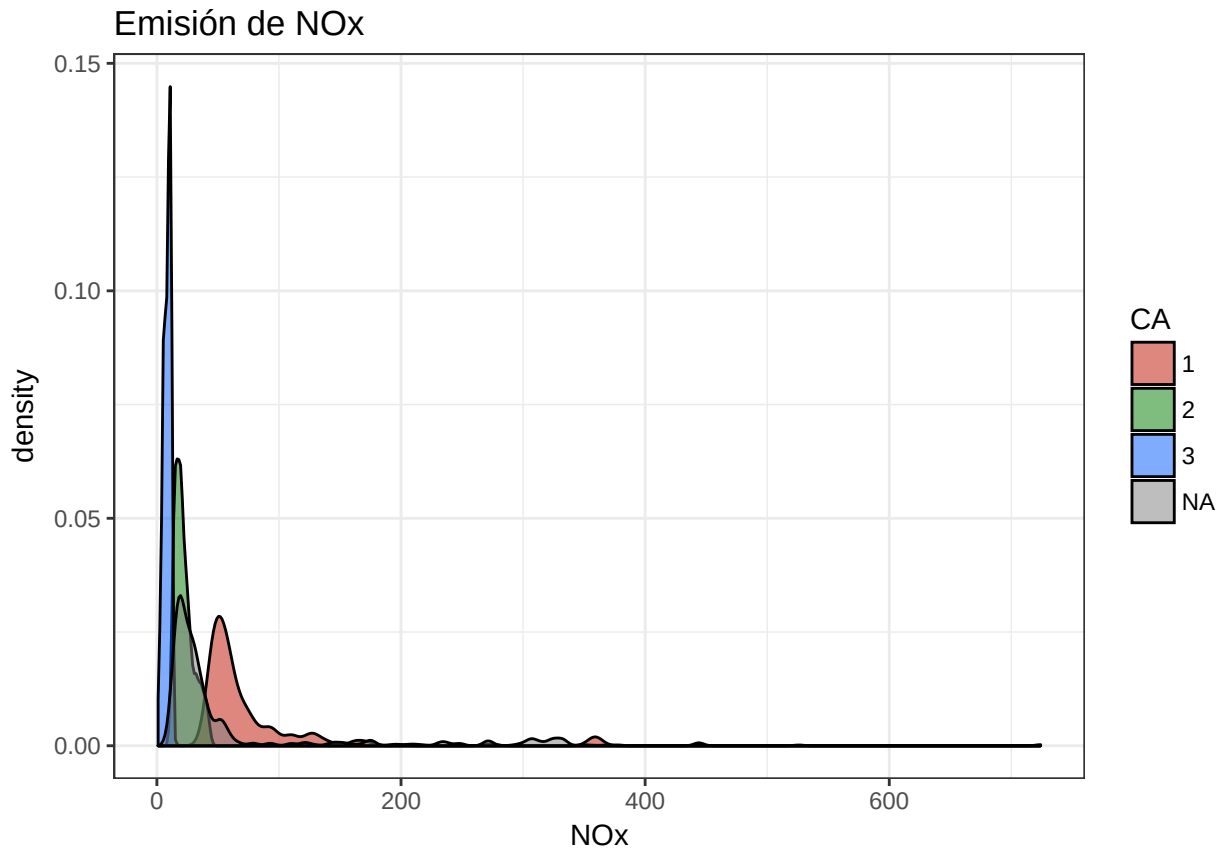
Para darnos una idea en cuanto a la manera en que se distribuye cada una de estas variables, en relación a cada una de las categorías representada por los niveles del atributo ***CA***, en lo que sigue se grafican sus funciones de densidad, por categoría de ***CA***

```
### Gráfica de densidad de CO2 sobre categorías del atributo CA
###
ggplot(data = cga2018, aes(x = CO2, fill = CA)) +
               geom_density(alpha = 0.5) +
               scale_fill_hue(l=40, c=135) +
               labs(title = "Emisión de CO2") +
               theme_bw()
```



Se aprecian formas de distribuciones bastante disparejas entre categorías de la variable de respuesta, quizá producto de la pronunciada diferencia en la cantidad de casos por clase; pero con una tendencia a concentrarse en un entorno del valor de los 200 gr/Km.

```
### Gráfica de densidad de NOx sobre categorias del atributo CA
###
ggplot(data = cga2018, aes(x = NOx, fill = CA)) +
               geom_density(alpha = 0.5) +
               scale_fill_hue(l=40, c=135) +
               labs(title = "Emisión de NOx") +
               theme_bw()
```

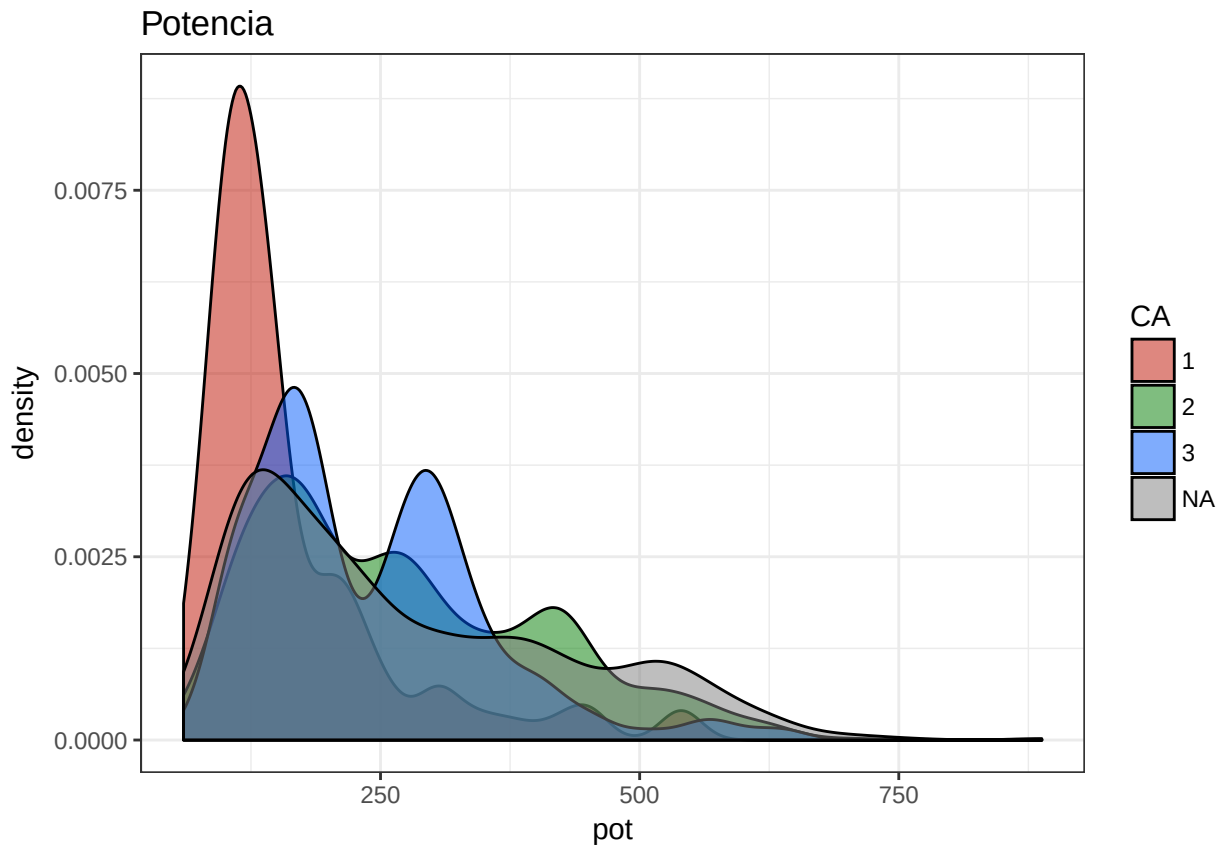


En esta otra variable, también vemos como varia la forma de la distribución por categoría de la variable de respuesta, pero asistimos a una propagación de los distintos grupos a lo largo del eje horizontal, que hace pensar en una posible diferencia de medias (importante en caso de aplicar algoritmos de clasificación).

Extendiendo la inspección gráfica al resto de las variables contínuas de interés, tenemos:

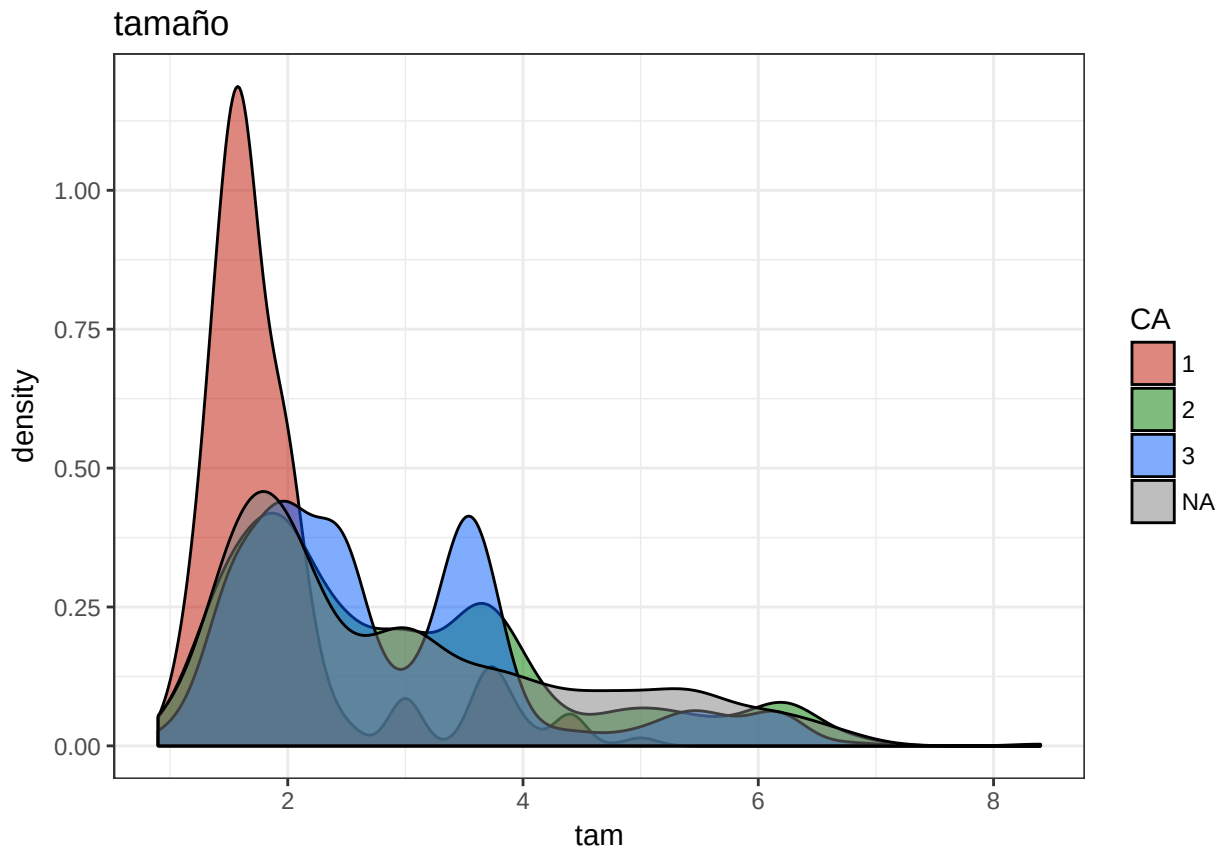
La variable ***potencia*** exhibe un comportamiento similar, con formas de distribución disparejas por categoría de respuesta, con una concentración de modelos de baja potencia.

```
### Gráfica de densidad de pot sobre categorías del atributo CA
###
ggplot(data = cga2018, aes(x = pot, fill = CA)) +
               geom_density(alpha = 0.5) +
               scale_fill_hue(l=40, c=135) +
               labs(title = "Potencia") +
               theme_bw()
```



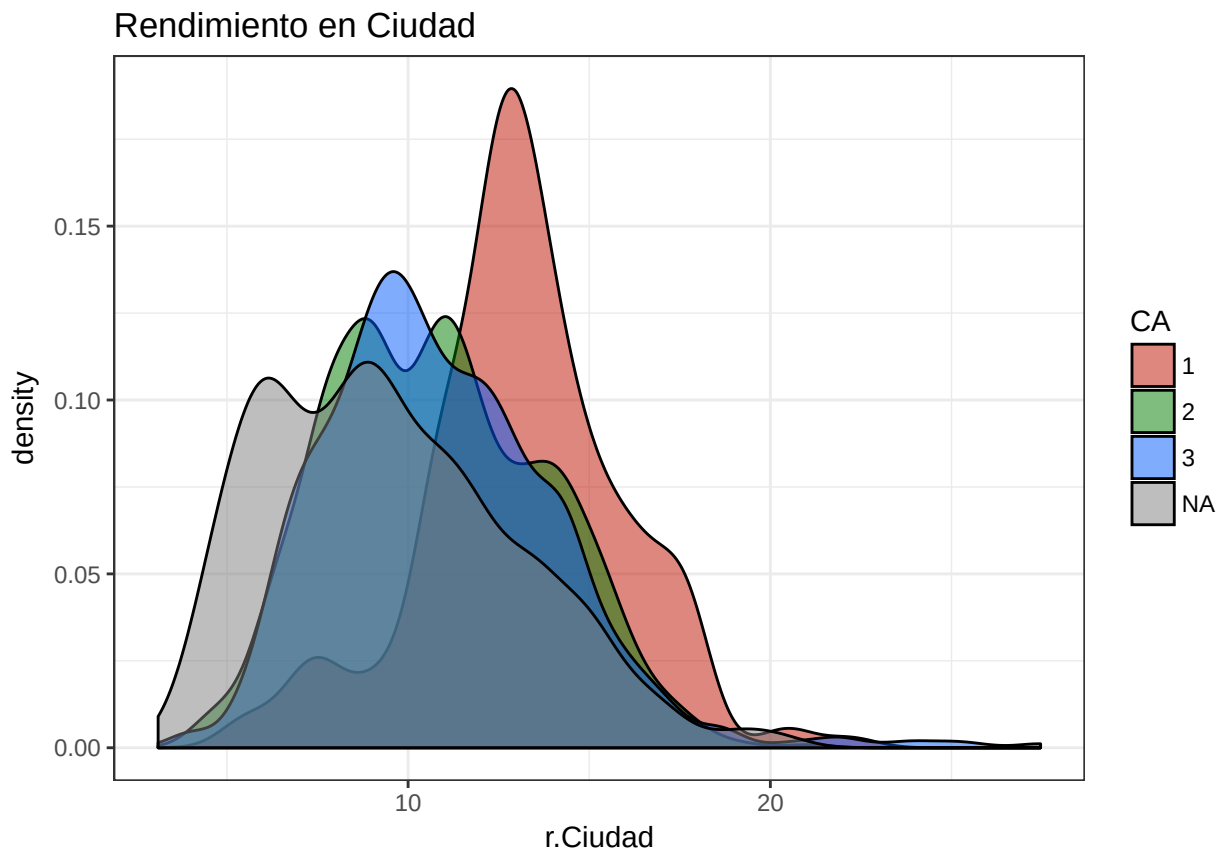
El atributo ***tam*** no es la excepción, y muestra un conducta similar a las anteriores, mostrando que la oferta de modelos está concentrada en motores con capacidad de alrededor de 2 litros.

```
### Gráfica de densidad de tam sobre categorías del atributo CA
###
ggplot(data = cga2018, aes(x = tam, fill = CA)) +
               geom_density(alpha = 0.5) +
               scale_fill_hue(l=40, c=135) +
               labs(title = "tamaño") +
               theme_bw()
```



Finalmente, para el ***rendimiento en Ciudad*** tenemos una situación muy paracida, con bastante más diferencia en las formas de las distribuciones parciales, con una aparente concentración alrededor de los 12 Km por litro.

```
### Gráfica de densidad de r.Ciudad sobre categorías del atributo CA
###
ggplot(data = cga2018, aes(x = r.Ciudad, fill = CA)) +
               geom_density(alpha = 0.5) +
               scale_fill_hue(l=40, c=135) +
               labs(title = "Rendimiento en Ciudad") +
               theme_bw()
```



Un aspecto particularmente llamativo es la manera en que los valores NA sobre la variable de respuesta ***CA*** tiende a cubrir algunos segmentos de los grupos graficados. La ausencia de observaciones completas en las proporciones observadas sobre éstos atributos, puede tener un efecto adverso en relación a la capacidad predictiva de dichas variables.

4.8 Selección de Atributos de la Muestra

La siguiente tabulación de niveles de ***Emisión de Gases de Efecto Invernadero***, por calificación de ***Contaminación de Aire*** nos permite apreciar mejor el efecto de la ausencia de valores (inicializados a cero) sobre los niveles de emisión ***GEI***, para la categoria **Baja** de ***CA***, que hace evidente un desbalance en la muestra.

```
##### Tabulación: Calificación de Contaminación de Aire vs Calificación de Gases de Efecto Invernadero
##
table(cga2018$CA, cga2018$GEI, dnn=c("Contaminación del Aire","Gases de Efecto Invernadero"))
```

```
##                       Gases de Efecto Invernadero
## Contaminación del Aire   1   2   3   4   5   6   7   8   9  10
##                      1   1   2  10   8  13  50  62  46   9   0
##                      2 112 111 181 189 183 209 267 181  23   0
##                      3 119  76 266 219 190 315 268 218  41  10
```

El total de valores tabulados por ***Emisión de Gases de Efecto Invernadero*** es de 3379 observaciones para la variable, que excluye las 1238 observaciones que corresponden a calificaciones de ***Contaminación de Aire*** con valores perdidos.

Sigue la extracción de observaciones para constituir el set de datos para la confección del modelo, orientado a

la estimación de la ***Contaminación del Aire***. Con ello se deja aparte al conjunto de observaciones que no reportan valores en este atributo, y que luego podrían ser utilizadas para la generación de predicciones al respecto (siempre que sea compatibles con las condiciones del modelo).

```
##### Extracción de filas correspondientes a observaciones completas
##
cga2018.dat <- cga2018[ which( is.na(cga2018$CA) != TRUE ), ]
nrow(cga2018.dat)
```

```
## [1] 3379
```

Producto de la reducción de filas, una revisión de las proporciones a través de una tabulación simple de variables categóricas del set, revela que la variable ***comb*** deja de tener una distribución por niveles (**Diesel**, **Gasolina**) que resulte relevante o razonable para su consideración

```
##### Tabulación de variables categóricas del set
##
table(cga2018.dat$trans,
      cga2018.dat$cil,
      cga2018.dat$comb,
      dnn=c("Tipo Transmisión","# Cilindros Invernadero","Tipo Combustible"))
```

```
## , , Tipo Combustible = Diesel
##
##                 # Cilindros Invernadero
## Tipo Transmisión   3   4   5   6   8  10  12
##     Auto           0   0   0   0   0   0   0
##     Dbl.Embrague   0   0   0   0   0   0   0
##     Manual         0   8   0   0   0   0   0
##     Manual.Robot   0   0   0   0   0   0   0
##     Var.Continua   0   0   0   0   0   0   0
##
## , , Tipo Combustible = Gasolina
##
##                 # Cilindros Invernadero
## Tipo Transmisión   3   4   5   6   8  10  12
##     Auto          25 710  34 571 332   0  40
##     Dbl.Embrague   0  96   3 173  16   0   0
##     Manual        17 620  22 158  32   9   0
##     Manual.Robot   3 131   7  52  21  10   0
##     Var.Continua   2 237   0  50   0   0   0
```

El tipo de combustible tendrá que ser erradicado del estudio, ya que al eliminar las observaciones referentes a valores perdidos del atributo ***Contaminación del Aire***, se pierde casi la totalidad de la representación del estrato **Diesel**.

Esta circunstancia está relacionada con la observación adelantada en la sección Distribución de Variables Continuas, en la que se menciona el solapamiento de observaciones con valores perdidos sobre la variable de respuesta ***CA***, sobre otras variables como la ***Potencia*** el ***Tamaño***.

Colateralmente, estas medidas de eliminación de observaciones, contribuyen a disminuir el efecto de la presencia de presuntos valores outlier en las columnas ***CO2*** y ***NOx***, producto a su vez de la presencia de motores **Diesel** en la muestra. Sin embargo, esta medida afecta sensiblemente la capacidad predictiva del modelo que se obtenga, ya que no estará en capacidad de predecir correctamente observaciones que correspondan a modelos que consuman este tipo de combustible.

Habiendo ajustado e imputado los valores de los atributos considerados como afectados por inconsistencias o por la presencia de valores perdidos, de seguido se suprimen los atributos que no constituyen características

técnicas del vehículo, sino que forman parte de su identificación o descripción.

En este sentido, junto con la columna ***comb*** se eliminan las columnas de carácter descriptivo: ***marca***, ***submarca***, ***ve*** (versión), ***modelo*** y ***categoría***.

```
##### Supresión de atributos Descriptivos:
##    Marca, Submarca, Versión, Modelo, Categoría y Combustible
##
cga2018.dat <- cga2018.dat[ which( cga2018.dat$comb != "Diesel" ), ]

cga2018.dat <- cga2018.dat[ c(-1, -2, -3, -4, -6, -10)]
str(cga2018.dat)
```

```
## 'data.frame':    3371 obs. of  12 variables:
##  $ trans      : Factor w/ 5 levels "Auto","Dbl.Embrague",..: 5 5 5 1 5 5 5 5 5 5 ...
##  $ cil        : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ pot        : num  188 188 188 188 156 156 90 90 90 110 ...
##  $ tam        : num  2 2 2 2 2.5 2.5 1.5 1.5 1.5 1.3 ...
##  $ r.Ciudad   : num  27.4 27.4 25.6 25.6 24 ...
##  $ r.Carretera: num  28.6 28.6 24.8 24.8 21.9 ...
##  $ r.Comb     : num  28.9 28.9 25.2 25.2 23 ...
##  $ r.Ajust    : num  21.7 21.7 18.9 18.9 17.3 ...
##  $ CO2        : num  107 107 123 123 135 135 119 119 127 134 ...
##  $ NOx        : num  5 5 2 2 5 6 7 7 7 10 ...
##  $ GEI        : Factor w/ 10 levels "1","2","3","4",..: 10 10 10 10 9 9 10 10 9 9 ...
##  $ CA         : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 3 3 3 3 ...
```

4.9 Normalización de Variables Numéricas

Como parte de las tareas de Limpieza y adecuación de datos, incluimos el escalamiento de las variables numéricas, de manera que en caso de aplicar métodos que impliquen cálculo de distancias, la diferencia de escalas no distorsione el resultado.

```
### Escalamiento de variables numéricas
###

normalizar <- function(data){ return (data - min(data))/(max(data) - min(data))}
normalizar_0 <- function(data){ return (data - mean(data))/(sd(data))}

cga2018.dat$pot <- normalizar(cga2018.dat$pot)
cga2018.dat$tam <- normalizar(cga2018.dat$tam)

cga2018.dat$r.Ciudad <- normalizar(cga2018.dat$r.Ciudad)
cga2018.dat$r.Comb   <- normalizar(cga2018.dat$r.Comb)
cga2018.dat$r.Ajust  <- normalizar(cga2018.dat$r.Ajust)
cga2018.dat$r.Carretera <- normalizar(cga2018.dat$r.Carretera)

cga2018.dat$CO2 <- normalizar(cga2018.dat$CO2)
cga2018.dat$NOx <- normalizar(cga2018.dat$NOx)
```

5. Análisis de los Datos.

5.1. Selección de Grupos a Analizar/Comparar

Siendo el propósito de la investigación, la confección de un modelo que permita **validar la calificación de nuevos vehículos en términos de su impacto ambiental** mediante la estimación de su potencial de **Contaminación de Aire**, la siguiente sección hará foco en extraer toda la información posible de la muestra, a efectos de orientar la selección y prueba de tipos de modelos, que resulten más adecuados a los efectos de cubrir el objetivo que nos hemos trazado.

Si bien, habiendo excluído el ***Tipo de Combustible***, las variables categóricas restantes ***Tipo de Transmisión***, ***Número de Cilindros*** y ***Emisión de Gases de Efecto Invernadero***, pueden ser de utilidad para la conformación de grupos que puedan ayudar a entender el comportamiento del modelo; resulta de mayor relevancia comprender el comportamiento de los datos en función a los niveles reportados de **CA**.

Por consiguiente, la caracterización más importante y referencia principal para la tarea de comparaciones y análisis estadísticos, estarán planteadas en función a la agrupación propuesta por la reducción aplicada sobre los Niveles de la variable de Respuesta ***CA***: **Baja**, **Moderada** y **Alta**.

Análisis a Aplicar.

- Verificación de la existencia de distribuciones normales en las variables continuas incluidas en la muestra. Aunque la impresión hasta ahora es que no hay columnas en la muestra que responda a una distribución normal, es necesario aplicar un método numérico para confirmarlo.
- Verificación de la condición de homogeneidad de varianza de los atributos continuos, para los grupos que resultan de la segmentación de la variable de respuesta.
- Verificación de la existencia de relaciones entre las columnas del set de datos, mediante la revisión de posibles correlaciones entre ellas. La idea es determinar condiciones de independencia, y evitar redundancias en la selección final de atributos.
- Validación por contraste de hipótesis de las Medias de las variables continuas en las que se aprecie homogeneidad sobre los distintos niveles de la variable de salida.

5.2. Comprobación de Normalidad y Homogeneidad de varianza.

5.2.1 Normalidad: Test Shapiro-Wilk

La condición de distribución normal de las variables tiene incidencia en las pruebas de inferencia y contraste de significancia de los parámetros de los distintos modelos. Sin embargo, el volumen de la muestra puede hacer menos sensibles los test a la ausencia de normalidad.

La apreciación obtenida a partir de las gráficas de cuantiles (qq), elaboradas para las variables numéricas en secciones previas, debe ser confirmada con los resultados derivados de métodos numéricos.

Para efectos del ejercicio, en el que tenemos una muestra razonable (menos de 5.000 observaciones), utilizaremos el método Shapiro-Wilk, que estipulada como hipótesis nula la normalidad de la muestra:

```
### Test de Normalidad Shapiro-Wilk para variables numéricas
###

nivel.significacion <- 0.05
col.nbr.dat <- colnames(cga2018.dat)
for(i in 1:ncol(cga2018.dat)){
   if (is.integer(cga2018.dat[ , i]) | is.numeric(cga2018.dat[ , i])) {
       pvlue <- shapiro.test(cga2018.dat[ , i])$p.value
```

```
      if (pvlue >= nivel.significacion) {
        print( sprintf("%11s : %8e *** ", col.nbr.dat[i], pvlue)  )
      }
      else {
        print( sprintf("%11s : %8e     ", col.nbr.dat[i], pvlue)  )
      }
    }
}
```

```
## [1] "        pot : 6.971368e-38     "
## [1] "        tam : 1.038833e-43     "
## [1] "   r.Ciudad : 9.037702e-24     "
## [1] "r.Carretera : 4.522249e-12     "
## [1] "     r.Comb : 2.521372e-17     "
## [1] "    r.Ajust : 2.521720e-17     "
## [1] "        CO2 : 4.337577e-38     "
## [1] "        NOx : 7.138961e-63     "
```

El listado de los p-valor resultantes, obtenidos para cada una de las columnas evaluadas, nos obliga a rechazar la hipótesis de normalidad para cada uno de los atributos enumerados, tal como se había anticipado al revisar las gráficas qq, en la sección dedicada a la descripción de los datos.

5.2.4 Homoscedasticidad: Tests de Levene y de Fligner

Se aplicará la prueba de Levene de Homogeneidad de Varianza, sobre las variables continuas ***pot***, ***Rendimiento en Ciudad*** y ***NOx***, que son las que hasta ahora parecen prometedoras en relación a su capacidad predictiva de la calificación de ***Contaminación de Aire***, considerando los grupos determinados por los niveles ***Baja***, ***Moderada***, y ***Alto***.

```
### Evaluación de Levene para Variables Continuas vs Variable de Respuesta
###
leveneTest(y=cga2018.dat$pot, group=cga2018.dat$CA, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    2  56.033 < 2.2e-16 ***
##       3368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y=cga2018.dat$r.Ciudad, group=cga2018.dat$CA, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value  Pr(>F)
## group    2  4.5573 0.01055 *
##       3368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y=cga2018.dat$NOx, group=cga2018.dat$CA, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    2  478.33 < 2.2e-16 ***
##       3368
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados obtenidos de la aplicación de la prueba sobre la variable ***pot***, resulta en un p-valor por debajo del nivel de significación de 0.05, que nos lleva a rechazar la hipótesis nula del test (de igualdad de varianza) de la variable, con respectos a los grupos definidos por la variable de respuesta.

Para las variables ***rendimiento en Ciudad*** y ***NOx*** observamos el mismo resultado.

La aplicación de la prueba de Fligner (no paramétrica, y no sensible a la ausencia de normalidad de los atributos) arroja resultados similares en cuanto a la heteroscedasticidad de la muestra:

```
### Evaluación de Filgner para Variables Continuas vs Variable de Respuesta
###
fligner.test(pot ~ CA, data= cga2018.dat)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  pot by CA
## Fligner-Killeen:med chi-squared = 146.28, df = 2, p-value <
## 2.2e-16
```

```
fligner.test(r.Ciudad ~ CA, data= cga2018.dat)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  r.Ciudad by CA
## Fligner-Killeen:med chi-squared = 7.9686, df = 2, p-value =
## 0.01861
```

```
fligner.test(NOx ~ CA, data= cga2018.dat)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  NOx by CA
## Fligner-Killeen:med chi-squared = 826.69, df = 2, p-value <
## 2.2e-16
```

En todos los casos, el p-valor obtenido en la prueba está por debajo del nivel de significación fijado.

En conclusión, las variables continuas analizadas tienen varianza distinta en al menos alguno de los niveles, en el dominio de la variable categórica utilizada para la definición de los grupos de interés.

Por lo tanto, para todos los casos probados (en ambas pruebas), se rechaza la hipótesis nula en beneficio de la alternativa, que establece la hetereoscedasticidad de las variables para cada uno de los grupos.

5.3. Aplicación de Pruebas Estadísticas.

Producto de la ausencia de normalidad en las variables de la muestra, y la ausencia homoscedasticidad en los grupos propuestos, no puede aplicarse el análisis de varianzas a través de el procedimiento anova, para validar la homogeneidad de las medias a través de los grupos del factor.

Nos restringiremos a revisar las relaciones entre columnas, y de las columnas versus la variable de respuesta, a través del estudio de la correlación que pueda existir entre los atributos; así como, a revisar la media de las variables continuas más importantes, con respecto a la variable de respuesta.

5.3.1 Variables Continuas Correlacionadas con la Respuesta

Con el resultado de la sección anterior, se confirma que las variables de la muestra no responden a una distribución normal. En consecuencia, para evaluar la correlación deberá optarse por el test de Spearman.

Sin embargo, la presencia de "tie values" impiden una correcta determinación de ρ (rho), por lo que se ensaya el método de Kendall como alternativa, explorando la determinación de τ_b (tau) como indicador de correlación:

```
### Exploración de la Correlación entre variables numéricas
###

cga2018.dat$CA <- as.integer(cga2018.dat$CA)

col.nbr.dat <- colnames(cga2018.dat)
hasta <- ncol(cga2018.dat) - 1 # supresión de variable respuesta

for(i in 1:hasta){
  if (is.integer(cga2018.dat[ , i]) | is.numeric(cga2018.dat[ , i])) {
    cor.rslt <- cor.test(cga2018.dat[ , i], cga2018.dat$CA, method="kendall")

    r <- cor.rslt$estimate
    pvlue <- cor.rslt$p.value

    if (pvlue >= nivel.significacion) {
      print( sprintf("%11s: p-vle=%8e  (tau=%8e)  (r^2=%8e) *** ", col.nbr.dat[i], pvlue, r, r*r)  )
    }
    else {
      print( sprintf("%11s: p-vle=%8e  (tau=%8e)  (r^2=%8e)    ", col.nbr.dat[i], pvlue, r, r*r)  )
    }
  }
}
```

```
## [1] "        pot: p-vle=6.754736e-01  (tau=5.776454e-03)  (r^2=3.336742e-05) *** "
## [1] "        tam: p-vle=2.221690e-06  (tau=6.698347e-02)  (r^2=4.486785e-03)    "
## [1] "   r.Ciudad: p-vle=2.921802e-06  (tau=-6.437585e-02)  (r^2=4.144250e-03)    "
## [1] "r.Carretera: p-vle=2.095483e-05  (tau=-5.855401e-02)  (r^2=3.428572e-03)    "
## [1] "     r.Comb: p-vle=6.402200e-03  (tau=-3.749069e-02)  (r^2=1.405551e-03)    "
## [1] "    r.Ajust: p-vle=6.355375e-03  (tau=-3.752832e-02)  (r^2=1.408374e-03)    "
## [1] "        CO2: p-vle=6.715185e-03  (tau=3.734126e-02)  (r^2=1.394369e-03)    "
## [1] "        NOx: p-vle=0.000000e+00  (tau=-7.558263e-01)  (r^2=5.712734e-01)    "
```

```
cga2018.dat$CA <- factor(cga2018.dat$CA)
```

La revisión del listado de resultado nos lleva al siguiente conclusión:

El atributo ***NOx*** exhibe un valor estimado para el parámetro τ_b de 0.7, para un coeficiente de correlación de 0.57, que confirma su relación con la variable de respuesta. Sin embargo, para el resto de las variables, el coeficiente obtenido es bastante próximo a cero, lo que indica ausencia de correlación.

5.3.2 Correlación Entre Variables Continuas

La correlación entre las variables de la muestra, permiten determinar relaciones que pueden derivar en la simplificación de los datos por eliminación de redundancias. En este sentido, se hace la inspección de las correlaciones entre columnas, apoyados en la graficación de la matriz de correlaciones, tal como se ilustra de seguido:

```
### Aproximación gráfica
###
aux <- data.frame(cga2018.dat$pot,
                  cga2018.dat$tam,
                  cga2018.dat$r.Ciudad,
                  cga2018.dat$r.Carretera,
                  cga2018.dat$r.Comb,
                  cga2018.dat$r.Ajust,
                  cga2018.dat$CO2,
                  cga2018.dat$NOx)

colnames(aux) <- c("Pot.", "Tam.",
                   "R.Cdad", "R.Carr.", "R.Comb.", "R.Ajust",
                   "CO2", "NOx")

M <-cor(aux)
corrplot.mixed(M, upper = "ellipse", number.cex = .7, tl.cex = .7)
```

La gráfica revela que hay correlaciones entre pares de atributos, que nos permitirán simplificar el tamaño de la muestra: Tanto la ***potencia*** como la ***capacidad*** del motor (tamaño), están positivamente correlacionadas con la emisión de ***CO2***.

Igualmente, el ***rendimiento en Ciudad***, ***Carretera***, por litro de ***Combustible*** y su ***Ajuste*** lo están entre sí, pero negativamente con los anteriores.

La variable correspondiente a la emisión de ***NOx***, si evidencia una conducta independiente.

5.3.3 Media de Atributos por Categoría de Respuesta

De la revisión de la correlación entre las variables numéricas, resulta la simplificación del conjunto inicial, para continuar con los atributos ***CO2***, ***NOx*** y ***Rendimiento en Ciudad***.

Media del atributo ***CO2*** por nivel del factor de salida se lista de seguido:

```
### Inspección de la Media por Categoría de Respuesta - atr CO2
###
for (i in 1:length(levels(cga2018.dat$CA))) {
  j <- levels(cga2018.dat$CA)[i]
  print( paste( "Contaminación Aire: categoría ",
                sprintf("(%s) %6.2f", j, mean(cga2018.dat$CO2[cga2018.dat$CA == j]) )) )
}
```

```
## [1] "Contaminación Aire: categoría  (1) 102.90"
## [1] "Contaminación Aire: categoría  (2) 149.71"
## [1] "Contaminación Aire: categoría  (3) 143.83"
```

Media del atributo ***NOx*** por nivel del factor de salida:

```
### Inspección de la Media por Categoría de Respuesta - atr NOx
###
for (i in 1:length(levels(cga2018.dat$CA))) {
  j <- levels(cga2018.dat$CA)[i]
  print(paste( "Contaminación Aire: categoría ",
                sprintf("(%s) %6.2f", j, mean(cga2018.dat$NOx[cga2018.dat$CA == j]) )) )
}
```

```
## [1] "Contaminación Aire: categoría  (1)  66.80"
## [1] "Contaminación Aire: categoría  (2)  21.21"
## [1] "Contaminación Aire: categoría  (3)   7.46"
```

Media del atributo ***rendimiento en Ciudad*** por nivel del factor de salida:

```
### Inspección de la Media por Categoría de Respuesta - atr r.Ciudad
###
for (i in 1:length(levels(cga2018.dat$CA))) {
  j <- levels(cga2018.dat$CA)[i]
  print(paste( "Contaminación Aire: categoría ",
                sprintf("(%s) %6.2f", j, mean(cga2018.dat$r.Ciudad[cga2018.dat$CA == j]) )) )
}
```

```
## [1] "Contaminación Aire: categoría  (1)   9.12"
## [1] "Contaminación Aire: categoría  (2)   6.81"
## [1] "Contaminación Aire: categoría  (3)   6.91"
```

De los resultados obtenidos en los distintos listados previos, los atributos ***CO2*** y ***NOx*** parecen exhibir buenas diferencias en sus medias, por nivel del factor de salida. No obstante, la variable ***rendimiento en Ciudad*** presenta valores bastante homogéneos.

Para este último caso es pruedente verificar si la variación en la media observada por categoría es estadísticamente significativa, y decidir si el campo posee suficiente información para incorporarlo en el modelo de clasificación.

Varia la media del rendimiento en Ciudad para el grupo de autos con CA de categoría Baja, con respecto al resto de la clasificación ?

Para responder a la pregunta se formarán los dos grupos, y se realizará un contraste de hipótesis para medias de dos muestras, especificando la condición de varianza desconocida y utilizando el nivel de significación del

0.05.

```
#### Comparación de la media del atributo r.Ciudad para los grupos superiores de CA
##   vs el grupo de calificación Baja
##
md.cdad.bajos <- subset(cga2018.dat, CA == 1, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 2 | CA == 3, select=c(r.Ciudad))

t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  md.cdad.bajos and md.cdad.altos
## t = 11.026, df = 223.04, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.853869 2.660765
## sample estimates:
## mean of x mean of y
##  9.124922  6.867605
```

El test se aplica en su la versión Welch (dos muestras con varianza desconocida), obteniendo un p-valor cercano a cero, que nos hace rechazar la hipótesis nula de igualdad de medias: Esto es, ambos grupos observan medias distintas.

Este resultado puede estar influido por la distribución de observaciones por categoría de ***CA*** que hemos ya identificado en secciones previas. Por lo que cabe preguntarse lo siguiente:

Para el Rendimiento en Ciudad, varían entre si las medias entre pares de categorías?

Para este tipo de preguntas pudiera aplicarse el análisis de varianza (anova), si las condiciones estuvieran dadas; pero para nuestra muestra (sin normalidad y heteroscedasticidad), tendremos que aplicar el contraste de medias para muestras distintas a pares.

Repitiendo el mismo ejercicio previo (a pares de categoría de ***CA***), se lista el conjunto de p-valores obtenido para la combinación de categorías ensayadas:

```
#### Comparación a pares de la media del atributo r.Ciudad para los grupos superiores
##
md.cdad.bajos <- subset(cga2018.dat, CA == 1, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 2, select=c(r.Ciudad))

cat.1.vs.2 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA == 2, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 3, select=c(r.Ciudad))

cat.2.vs.3 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA == 1, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 3, select=c(r.Ciudad))

cat.1.vs.3 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

rslt <- cbind(cat.1.vs.2$p.value, cat.2.vs.3$p.value, cat.1.vs.3$p.value)
colnames(rslt) <- c("1.vs.2", "2.vs.3", "1.vs.3")
rslt
```

```
##            1.vs.2   2.vs.3       1.vs.3
## [1,] 5.358742e-23 0.361313 1.823994e-21
```

Si bien el valor-p para las comparaciones de las categorias **Moderada** (2) y **Alta** (3) en relación a la categoría **Baja** (1), habría que rechazar la hipótesis nula de igualdad de medias, el valor obtenido en la comparación entre las categorías (2) y (3), resulta por encima del nivel de significación, por lo que en este caso no puede rechazarse la igualdad de media entre.

Por lo anterior, el atributo ***rendimiento en Ciudad*** podría no tener suficiente capacidad predictiva para diferenciar vehiculos entre niveles de ***Contaminación de Aire***.

Varia la media del CO2 para el grupo de autos con CA para la categoría 1, con respecto al resto (2 y 3), para el subgrupo de autos de cuatro cilindros ?

Siendo siempre la intención la de evaluar la capacidad de discriminación del atributo versus la variable de respuesta, la pregunta va en línea con la intención original de utilizar las variables categóricas seleccionadas de la muestra para la determinar si existe alguna interacción adicional entre grupos, que pueda resultar de interés.

Nuevamente, la imposibilidad de aplicar la técnica del anova nos restringe a procedimientos menos poderosos como el previo, y análisis más limitados.

Al igual que el caso del ***rendimiento en Ciudad*** se toma los grupos categorías de ***CA***, pero esta vez restringiéndolos por el ***Número de Cilindros***, para finalmente aplicar el test:

```
#### Comparación de la media del atributo CO2 para los grupos superiores de CA (2,3)
##   versus los de categoría (1), en coches de 4 cilindros
##
md.cdad.bajos <- subset(cga2018.dat, (CA==1)       & (cil==4),select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, (CA==2|CA==3) & (cil==4),select=c(CO2))

t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  md.cdad.bajos and md.cdad.altos
## t = -5.1591, df = 236.88, p-value = 5.245e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.929671  -7.125793
## sample estimates:
## mean of x mean of y
##   89.2284  100.7561
```

Al igual que en el ejemplo previo, el test obliga a rechazar la hipótesis de igualdad de medias para ambos grupos (categorías **Bajas** de ***CA***, versus la agrupación de las **Moderadas** y **Altas**).

Al igual que antes, si bien para la categoría de Contaminación de Aire ***CA*** **Baja** se observa una media en relación a la emisión de ***CO2*** distinta a la que se tiene para el grupo combinado de categorías **Moderada** y **Alta**, debemos preguntarnos si entre la **Moderada** y la **Alta** se preserva la condición de medias distintas.

Dentro del grupo que concentra más ocurrencia de observaciones (grupo de categorías altas) serán distintas las medias del atributo CO2 entre las categorías ?

```
#### Comparación a pares de la media del atributo CO2 para los grupos superiores
##   en autos de cuatro cilindros
##
md.cdad.bajos <- subset(cga2018.dat, CA==2 & cil==4, select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, CA==3 & cil==4, select=c(CO2))
```

```r
t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  md.cdad.bajos and md.cdad.altos
## t = -0.80741, df = 1488.8, p-value = 0.4196
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.276871  2.199492
## sample estimates:
## mean of x mean of y
##  99.87176 101.41045
```

Se obtiene un p-valor que permite afirmar que no hay evidencia de que las medias de estas categorias sean significativamente distintas.

Serán distintas las medias del atributo NOx entre las categorías CA, para coches con transmisión automática ?

El poco poder predictivo que intuímos para los atributos ***rendimiento Ciudad*** y ***CO2***, producto de las pruebas anteriores, nos conduce a cuestionarnos el potencial de la variable ***NOx***. Por ella razón planteamos esta última pregunta, y aplicando el mismo enfoque previo:

```r
#### Comparación a pares de la media del atributo NOx
##   en autos de transmisión automática
##
md.cdad.bajos <- subset(cga2018.dat, CA==1 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==2 & trans=="Auto", select=c(NOx))

cat.1.vs.2 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==2 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==3 & trans=="Auto", select=c(NOx))

cat.2.vs.3 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==1 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==3 & trans=="Auto", select=c(NOx))

cat.1.vs.3 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

rslt <- cbind(cat.1.vs.2$p.value,
              cat.2.vs.3$p.value,
              cat.1.vs.3$p.value)

colnames(rslt) <- c("1.vs.2", "2.vs.3", "1.vs.3")
rslt
```

```
##            1.vs.2       2.vs.3       1.vs.3
## [1,] 2.219164e-20 8.057156e-215 2.756594e-26
```

Los valores obtenidos para el p-valor en todas las comparaciones están por debajo del valor de significación, por lo que las medias son significativamente distintas.

Conclusión de Sección: La única variable continua que presenta valores distintos de la media para los distintos niveles del factor ***Contaminación del Aire*** , es el atributo ***NOx***. El resto de las variables, no

cuenta con suficiente capacidad predictiva para distinguir observaciones que pertenecen a las categorías **Moderada** y/o **Alta**.

6. Modelo Tentativo

En el siguiente segmento se procede a extraer una parte del conjunto de datos para utilizarlo como set de pruebas, durante la evaluación de la precisión de los modelos que sean generados.

```
#### Formación del set de datos de Entrenamiento y de Test

sample.size <- nrow(cga2018.dat)     # Total de observaciones cargadas
m.training.size <- 0.8               # Porcentaje de la Muestra dedicada a Entrenamiento
set.seed(1965)

### Muestreo aleatorio para extraer el set de Entrenamiento
###
m.training.set <- sample( sample.size, sample.size * m.training.size )

# Tomando los valores esperados de la variable de respuesta para el set de prueba
#
m.test.rsl <- cga2018.dat[ -m.training.set, "CA" ]

# Se separa el set de entrenamiento del Arbol
#
m.training.tree <- cga2018.dat[ m.training.set, ]

## Eliminación de atributos redundantes o que han demostrado baja contribución
## por su escaso poder predictivo, además de la variable objetivo
##
cga2018.dat <- cga2018.dat[ , c( -3, -4, -5, -6, -7, -8, -12)]

# Se separa el set de entrenamiento del kmeans (kproto)
#
m.training <- cga2018.dat[ m.training.set, ]
m.test <- cga2018.dat[ -m.training.set, ]
```

6.1 Ensayo: Modelo de Agrupación

Considerando que las variables que tienden a tener mayor fuerza predictiva están disponibles para todos los tipos de vehículos listados en la data original, una primera aproximación interesante es la de considerar un algoritmo de agrupación, que en caso de resultar pudiera incluir en sucesivos experimentos las filas eliminadas a causa de la supresión de la variable ***Tipo de Combustible*** (excluida del ejercicio por falta de representatividad en la muestra, para resultados de la variable de respuesta).

Un método no supervisado como el **k-means** pudiera ser de utilidad, pero en su versión para datos mixtos (numéricos y categóricos), cuya implementación está disponible a través de la función **kproto**.

En vista de que hemos rexpresado en 3 niveles al atributo ***CA***, impondremos k = 3 para la inicialización del hiperparámetro del modelo.

```
#### kmeans para data mixta: numérica y categórica
##
mv <- kproto(m.training, 3)
```

```
## # NAs in variables:
## trans   cil   CO2   NOx   GEI
##     0     0     0     0     0
## 0 observation(s) with NAs.
##
## Estimated lambda: 3923.741
```

```
#clprofiles(mv, m.training)

m.predicted.clusters <- predict(mv, m.test)
m.pred <- factor(m.predicted.clusters$cluster)

confusionMatrix(m.pred, m.test.rsl)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3
##          1   1 130 139
##          2  20  65  97
##          3  13  91 119
##
## Overall Statistics
##
##                Accuracy : 0.2741
##                  95% CI : (0.2407, 0.3094)
##     No Information Rate : 0.5259
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0492
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity          0.029412   0.2273   0.3352
## Specificity          0.580343   0.6992   0.6750
## Pos Pred Value       0.003704   0.3571   0.5336
## Neg Pred Value       0.918519   0.5517   0.4779
## Prevalence           0.050370   0.4237   0.5259
## Detection Rate       0.001481   0.0963   0.1763
## Detection Prevalence 0.400000   0.2696   0.3304
## Balanced Accuracy    0.304877   0.4633   0.5051
```

La precisión obtenida es demasiado baja como para promover este modelo, que en parte nos reafirma el caracter poco predictivo de los datos.

6.2 Ensayo: Modelo de Clasificación

Alternativamente, consideraremos un modelo de árbol que ayude a ilustrar mejor la realidad de las caracteríticas de nuestro set de datos. Para ello utilizaremos las variables contínuas ***NOx*** y ***CO2***, además de las variables categporicas que hemos analizado en apartes anteriores (***trans***, ***cil*** y ***GEI***):

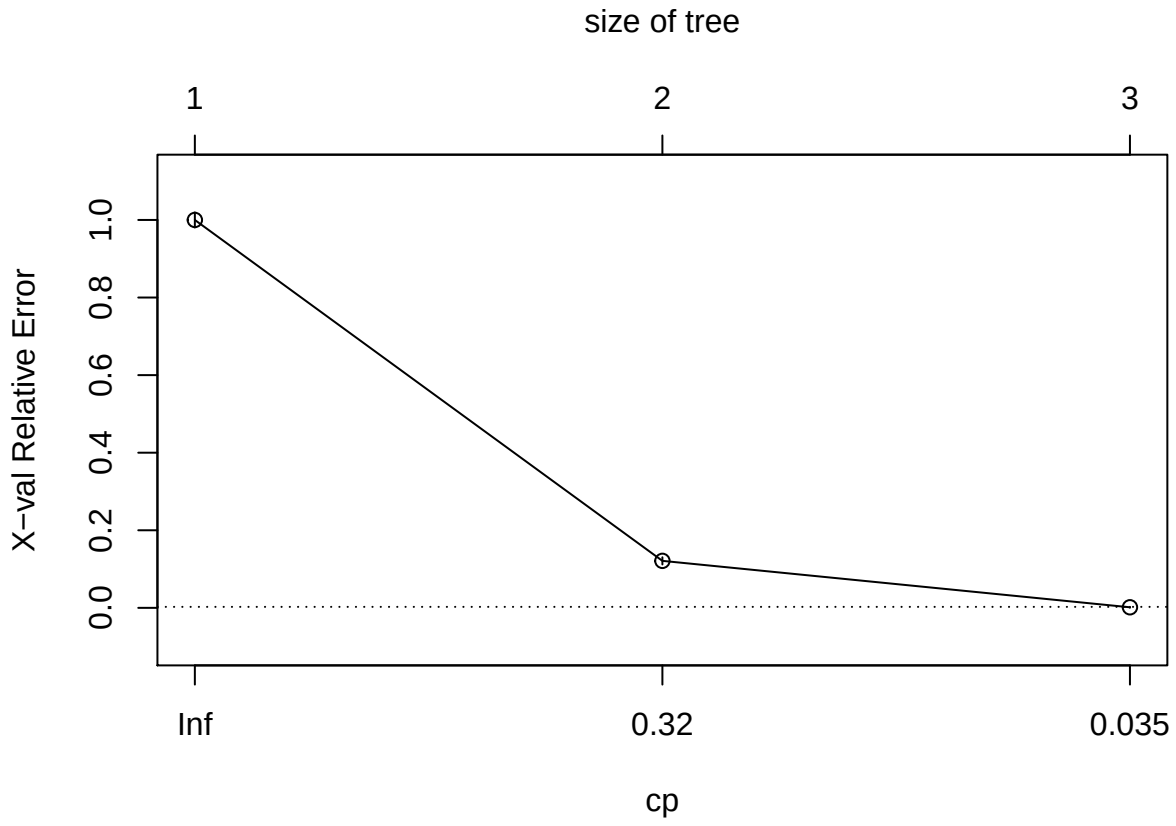
```
#### Modelo de árbol CART
##
```

```r
mv <- rpart(CA~NOx+CO2+trans+cil+GEI, data=m.training.tree, method="class")

printcp(mv) # resultados del modelo
```

```
##
## Classification tree:
## rpart(formula = CA ~ NOx + CO2 + trans + cil + GEI, data = m.training.tree,
##     method = "class")
##
## Variables actually used in tree construction:
## [1] NOx
##
## Root node error: 1329/2696 = 0.49295
##
## n= 2696
##
##         CP nsplit rel error    xerror      xstd
## 1 0.87886      0 1.0000000 1.0000000 0.0195327
## 2 0.11964      1 0.1211437 0.1211437 0.0092580
## 3 0.01000      2 0.0015049 0.0015049 0.0010637
```

```r
plotcp(mv)  # visualización de resultados
```

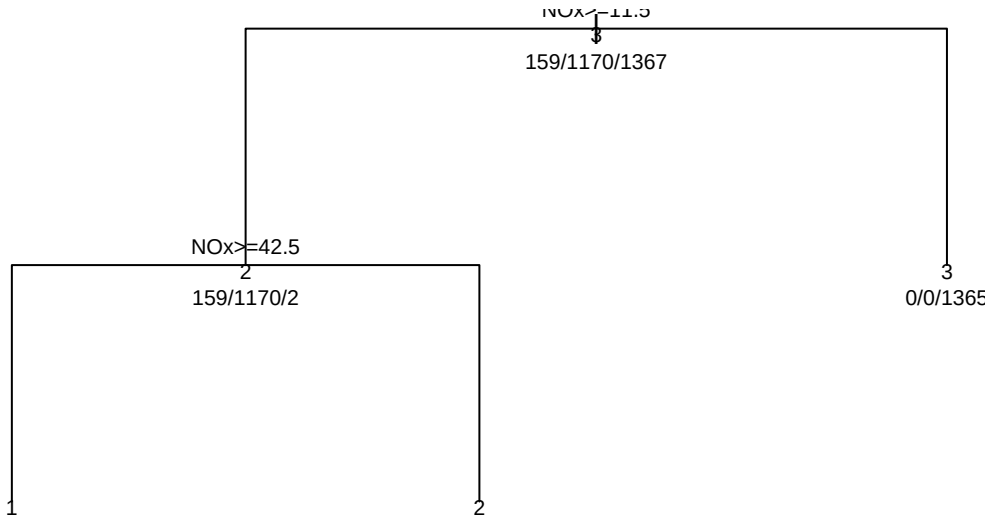


El algoritmo considera al atributo ***NOx*** suficiente para la confección del árbol, cuestión que intuitivamente se anticipaba a partir del estudio de la media, realizado en secciones previas.

El árbol es bastante pequeño, y el nivel de complejidad que minimiza el error es que el corresponde al CP 2,

tal como se ve en la tabulación de resultados y gráfica correspondiente.

```
#### Modelo de árbol CART
##
plot(mv, uniform=TRUE, main="Clasificación de CA")
text(mv, use.n=TRUE, all=TRUE, cex=.7)
```



Como parte del resultado de la construcción del árbol, se obtiene la calificación de importancia que da el algoritmo a los atributos considerados para la confección. La información se extrae de la variable variable.importance, obtenida del objeto summary(), resultante de aplicar la función al modelo.

```
##        NOx      trans        GEI        CO2        cil
## 1481.74347  168.00209   84.00104   60.51688   44.25861
```

Utilizando el conjunto reservado para el test, y estimando la precisión del modelo:

```
#### Evaluación del MOdelo de árbol CART
##

##  Obtención de predicción a partir del set de Prueba
###
m.rpart.predict <- predict(object = mv,
                                  newdata = m.test,
                                  type = "class")
m.pred <- factor(m.rpart.predict)
#
# Matríz de confusión construida con los resultados de la predicción vs
# la clasificación disponible en el set de prueba ...
#
```

```
confusionMatrix(m.pred, m.test.rsl)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3
##          1  34   0   0
##          2   0 286   0
##          3   0   0 355
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9945, 1)
##     No Information Rate : 0.5259
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity           1.00000   1.0000   1.0000
## Specificity           1.00000   1.0000   1.0000
## Pos Pred Value        1.00000   1.0000   1.0000
## Neg Pred Value        1.00000   1.0000   1.0000
## Prevalence            0.05037   0.4237   0.5259
## Detection Rate        0.05037   0.4237   0.5259
## Detection Prevalence  0.05037   0.4237   0.5259
## Balanced Accuracy     1.00000   1.0000   1.0000
```

La tasa de precisión que alcanza este nuevo modelo basado en un árbol CART, lo señala como muy adecuado el problema que nos hemos planteado.

6.3 Ensayo: Modelo de Regresión Logística

Si consideremos que la proporción de vehículos asociados a la categoría de contaminación del aire **Baja** es realmente muy baja en comparación con las otras dos, puede proponorse su integración a la categoría **Moderada**, de manera de convertir el problema de clasificación que hemos manejado hasta ahora, en uno de tipo binario.

Bajo el cancepto anterior, se desea evaluar la probabilidad de que un vehículo correspondiente a una nueva observación pueda ser clasificado como perteneciente a la categoría **Moderada** o **Alta**, para ello puede utilizarse un modelo regresión logística, estimado con las variables estudiadas hasta ahora. Particularmente, tomaremos como referencia la importancia con la que el algoritmo CART califica a las variables del set.

La variable dependiente será una variable binaria que indicará si el vehículo pertenece a la categoria de Contaminación de Aire **Alta**, o no

```
### Se incorpora variable dicotomica con 1 para vehículos con CA Alta, 0 al resto
###
m.training.tree$high <- ifelse(m.training.tree$CA == 3, 1, 0)
m.training.tree$high <- factor(m.training.tree$high)
```

```
m.test.rsl <- ifelse(m.test.rsl == 3, 1, 0)
m.test.rsl <- factor(m.test.rsl)
```

Se estima el modelo a través de una regresión logística, utilizando solo las variables relacionadas con emisiones ***NOx*** y ***CO2***, en vista de que las restantes ya han probado su escasa capacidad predictiva:

```
## Estimación del modelo de Regresión Logística para determinar la
## probabilidad de estar en la categoria CA Alta
## (solo con atributos de emision NOx y CO2)
#
mv <- glm(high ~ NOx+CO2, m.training.tree, family=binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mv)
```

```
##
## Call:
## glm(formula = high ~ NOx + CO2, family = binomial(), data = m.training.tree)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4642   0.0000   0.0000   0.0000   6.3281
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 68.8024536  6.5073459  10.573   <2e-16 ***
## NOx         -5.9252370  0.5544452 -10.687   <2e-16 ***
## CO2          0.0004602  0.0043342   0.106    0.915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3736.91  on 2695  degrees of freedom
## Residual deviance:  108.94  on 2693  degrees of freedom
## AIC: 114.94
##
## Number of Fisher Scoring iterations: 13
```

Tal como ya sabiamos de secciones anteriores, el sumario del modelo nos confirma que si bien el atributo ***NOx*** es significativo para la regresión, la variable ***CO2*** no permite rechazar la hiptesis nula, por lo que el parámetro no es significativo para el modelo.

Refactorizando solo con el atributo ***NOx*** en la ecuación:

```
## Estimación del modelo de Regresión Logística para determinar la
## probabilidad de estar en la categoria CA Alta
#
mv <- glm(high ~ NOx, m.training.tree, family=binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mv)
```

```
##
## Call:
## glm(formula = high ~ NOx, family = binomial(), data = m.training.tree)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4517   0.0000   0.0000   0.0000   6.3286
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  68.9447     6.3860   10.80   <2e-16 ***
## NOx          -5.9313     0.5522  -10.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3736.91  on 2695  degrees of freedom
## Residual deviance:  108.95  on 2694  degrees of freedom
## AIC: 112.95
##
## Number of Fisher Scoring iterations: 13
```

El valor del indicador AIC se muestra ligeramente inferior al previamente obtenido, por lo que se trata este último de un mejor modelo.

Para evaluar el modelo, determinaremos la Matriz de confusión, suponiendo un umbral de discriminación del 95%.

```
### Predicciones del modelo a partir de la data de entrenamiento
###
m.pred <- predict(mv, newdata = m.training.tree, type = "response")

### Confección de Matriz de Confusión al 95%
### utilizando como referencia la variable binaria "high"" introducida para la
### estimación del modelo
###
lvs <- c("Alta", "Otras")

highCA   <- factor(m.training.tree$high , labels = rev(lvs))
highCA.p <- factor(as.numeric(m.pred>0.95), labels = rev(lvs))

caret::confusionMatrix(highCA.p, highCA, positive = "Alta")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Otras Alta
##      Otras  1329    2
##      Alta      0 1365
##
##                Accuracy : 0.9993
##                  95% CI : (0.9973, 0.9999)
##     No Information Rate : 0.507
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9985
##  Mcnemar's Test P-Value : 0.4795
##
```

```
##             Sensitivity : 0.9985
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9985
##              Prevalence : 0.5070
##          Detection Rate : 0.5063
##    Detection Prevalence : 0.5063
##       Balanced Accuracy : 0.9993
##
##        'Positive' Class : Alta
##
```

El resultado es bastante bueno, con una precisión tan alta como la que muestra el árbol de la sección anterior, siendo el modelo capaz de reconocer todos los vehículos de categorías inferiores (Specificity con valor 1).

7. Conclusiones

A lo largo del trabajo se ha tomado un juego de datos público, y se le han aplicado una serie de análisis y técnicas de limpieza y preparación, con el objetivo de construir un modelo analítico que facilite la clasificación de los casos.

Así, se han identificado y resuelto casos de valores perdidos, y agrupaciones de valores en categoría más altas (con la intención de simplificar la dimensión del problema). Ante la presencia de valores de tipo outliers para las variables de la muestra, se optó por no ajustarlas, por tratarse de especificaciones técnicas que pudieran estar respondiendo a características legítimas de la oferta, justificadas por la alta segmentación que experimenta el mercado automotriz.

Específicamente, cualquier intento de tratamiento de los casos para los valores reportados en relación a las emisiones (presuntos outliers), solo puede ser procesadas bajo conocimiento funcional del dominio del problema, de modo que las acciones que se puedan adelantar en este sentido, sean coherentes y no desvirtúen la calidad de los datos. No obstante, una breve investigación sobre el tema, nos hacen pensar que los niveles reportados, son correctos y son similares a mediciones practicadas por instituciones como la ADAC.

Esta última observación es particularmente importante, ya que en parte, la normalidad de las variables pudo haber estado alterada por la presencia de tales valores.

Se han evaluado las distintas variables incluidas en el set, y la forma en que se distribuyen en relación a la variable objetivo que se ha definido, para finalmente hacer una selección en base los resultados obtenidos por las pruebas de correlación y el potencial predictivo que les es propio (a través del estudio del comportamiento de las medias por categoría respuesta).

En general, la ausencia de normalidad en las variables continuas de la muestra, así como la presencia de hetereoscedasticidad, nos llevó a utilizar versiones no paramétricas de los test estadísticos. Particularmente, resultó una limitante importante durante el análisis estadístico, en vista de la imposibilidad de utilizar la técnica anova para el estudio de las medias en función a los grupos establecidos por las variables categóricas.

Lamentablemente el set de datos presenta un déficit estructural con respecto a la casuística que debería considerarse dentro del dominio del problema, ya que no se dispone de suficientes valores para uno de los estratos de una de las variables, además de que para algunas categorías de la variable de respuesta se aprecia un desbalance en lo referente a la disponibilidad de casos etiquetados para algunas de las categorías:

`Particualarmente, el modelo de árbol para la calificación del nivel de la Contaminación del Aire que hemos obtenido, para la predicción del nivel de contaminación generado por vehículos a Gasolina, no puede ser aplicado a observaciones correspondientes a modelos que consumen Diesel como combustible.`

El problema que nos hemos planteado es de tipo Clasificación, tal como se recoge en el resultado obtenido, su comparación con un intento previo de aplicar un método de agrupamiento, nos confirma que otros tipos de modelos pueden resultar inútiles e inaplicables a causa de la escasa precisión que obtienen por la mezcla de tipos de datos que presenta la muestra, y la estructura misma de la información que engloban.

El desarrollo de un último modelo de Regresión Logística ha resultado muy acertivo, pero en un escenario en el que se reducen aún más las categorías de contaminación de Aire. Si bien, no es tan bueno como el de árbol, puede ser utilizado como un modelo alternativo en caso de que se ajuste a las restricciones de uso, pero debe ternerse en cuenta que sigue siendo válida la restricción de uso sobre vehículos a **Diesel**, ya que el modelo no ha sido entrenado con datos pertenecientes a este estrato.