

# Integración, Limpieza, Validación y Análisis de Datos.

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

*Antonio Carcia*

*diciembre 24, 2018*

## Contents

<b>Sumario</b>	<b>1</b>
<b>1. Introducción</b>	<b>2</b>
Objetivos: . . . . .	2
<b>2. Descripción del dataset.</b>	<b>2</b>
Propósito: . . . . .	3
<b>3. Integración y selección de los datos de interés a analizar.</b>	<b>3</b>
<b>4. Limpieza de los datos.</b>	<b>5</b>
4.1. Elementos vacíos, ceros y Nulos . . . . .	5
4.2. Identificación y tratamiento de valores extremos . . . . .	11
4.3 Exportación de Datos Procesados . . . . .	14
4.4 Distribución de la Variable de Respuesta . . . . .	14
4.5 Distribución de Variables Categóricas . . . . .	15
4.6 Distribución de Variables Continuas . . . . .	20
4.7 Selección de Atributos de la Muestra . . . . .	25
4.8 Normalización de Variables Numéricas . . . . .	27
<b>5. Análisis de los datos.</b>	<b>27</b>
5.1. Selección de los grupos de datos a analizar/comparar . . . . .	27
5.2. Comprobación de la normalidad y homogeneidad de la varianza. . . . .	28
5.3. Aplicación de pruebas estadísticas. . . . .	31
6. Modelo Tentativo . . . . .	38
7. Conclusiones . . . . .	44

## Sumario

**Fuente:** [www.kaggle.com](http://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
<b>Marca</b>	Fabricante del vehículo
<b>Submarca</b>	Modelo de comercialización
<b>Version</b>	Versión de fabricación - breve descripción de características
<b>Modelo</b>	Año al que corresponde el modelo
<b>Trans</b>	Modelo de Transmisión específico o Tipo utilizado en la fabricación del vehículo
<b>Comb</b>	Tipo de Carburante o energía utilizada por el Motor del Vehículo
<b>Cilindros</b>	Cantidad de Cilindros que configuran el Motor
<b>Potencia (HP)</b>	Potencia del Motor
<b>Tamaño (L)</b>	Capacidad del Motor
<b>Categoría</b>	Estilo o Clase de coche
<b>R. Ciudad (Km/L)</b>	Recorrido en Ciudad medido en km por litro de combustible
<b>R. Carr (Km/L)</b>	Recorrido en Carretera medido en km por litro de combustible

Aspecto	Descripción
<i>R. Comb (Km/L)</i>	Recorrido en Km por Litro
<i>R. Ajust (KM/L)</i>	Recorrido Ajustado medido en Km por Litro
<i>CO<sub>2</sub> (gr/Km)</i>	Emisiones de dióxido de carbono en gramos por Km recorrido
<i>NO<sub>x</sub> (gr/1000Km)</i>	Emisiones de óxido nitroso en gramos por unidades de 1000 Km
<i>Calificación Gas Ef. Inv.</i>	Calificación obtenida en medición de Gases de Efecto Invernadero emitidos
<i>Calificación Contam. Aire</i>	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 es 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 recogida en los datos utilizados como insumo, 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 a analizar.

El set de datos se asume autocontenido, por lo que no es necesario enriquecerlo o 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 COMPACTOS" ...
## $ 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 vacíos. 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",          # Categoría del Vehículo
```

```

        "r.Ciudad",      # Recorrido en Ciudad (Rendimiento por Litro)
        "r.Carretera",   # Recorrido en Carretera (Rendimiento por Litro)
        "r.Comb",        # Recorrido por Combustible (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 sustituye los placeholders que indican valores perdidos sobre la variable (carácter '?') 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 los 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 inicializados a Cero para atributo
###
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 análisis al atributo **GEI**, aplicando el mismo razonamiento 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 se constituyen una contradicción.

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 DOBLE TURBO 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 resultados 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*, inicilizados 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 )

```

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"

```

## 4.2. Identificación y tratamiento de valores extremos

A título de referencia se describe cuantitativamente la muestra.

```

### Explorando algunas estadísticas asociadas a los atributos de la muestra
###

```

```
summary(cga2018$pot)
```

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

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

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

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##      3.10      8.20     10.42     10.60     12.81     27.46
```

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

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

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

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

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

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

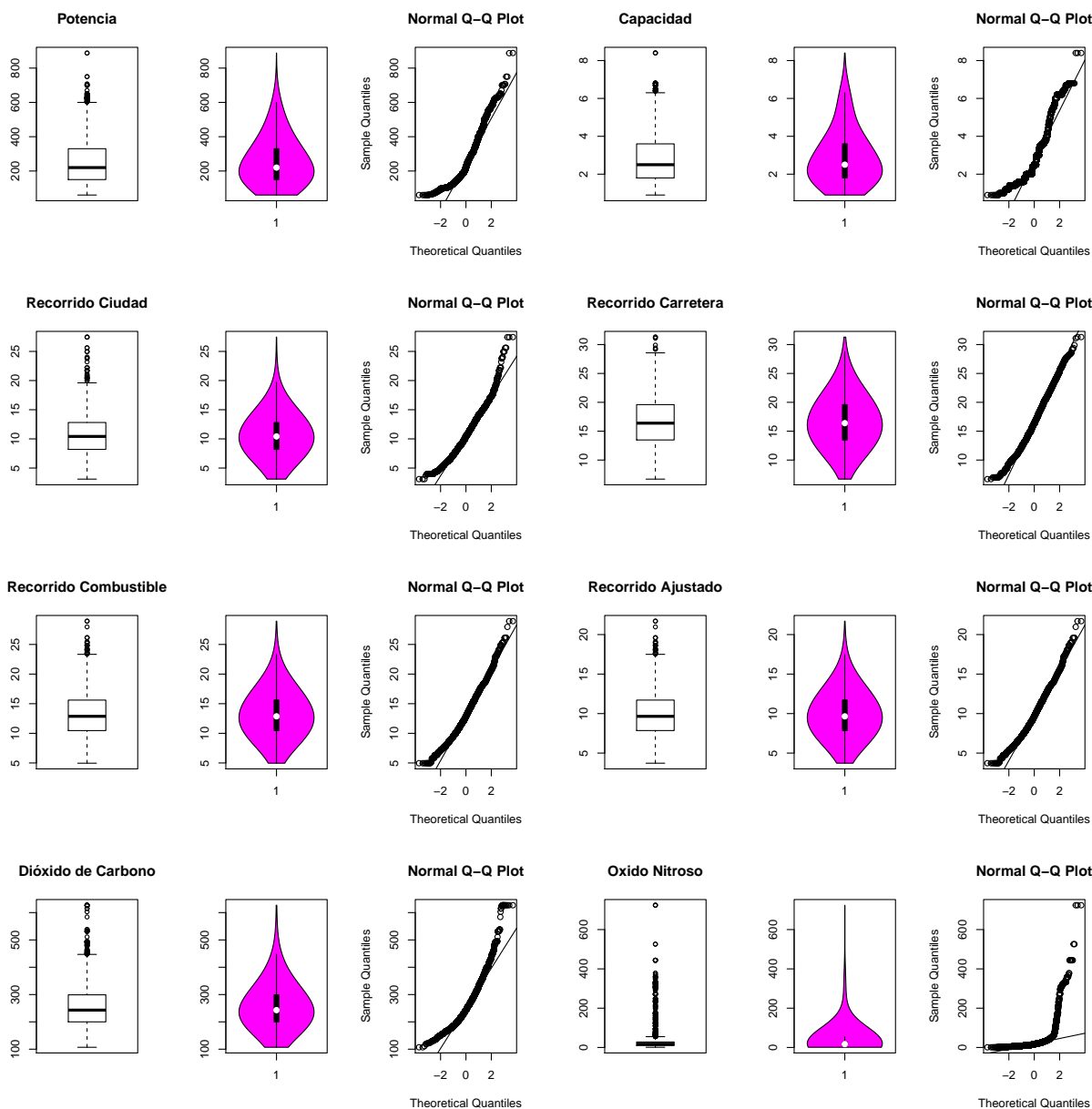
```
summary(cga2018$C02)
```

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

```
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 han 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 repercute 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 reportan en promedio 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*.

### 4.3 Exportación de Datos Procesados

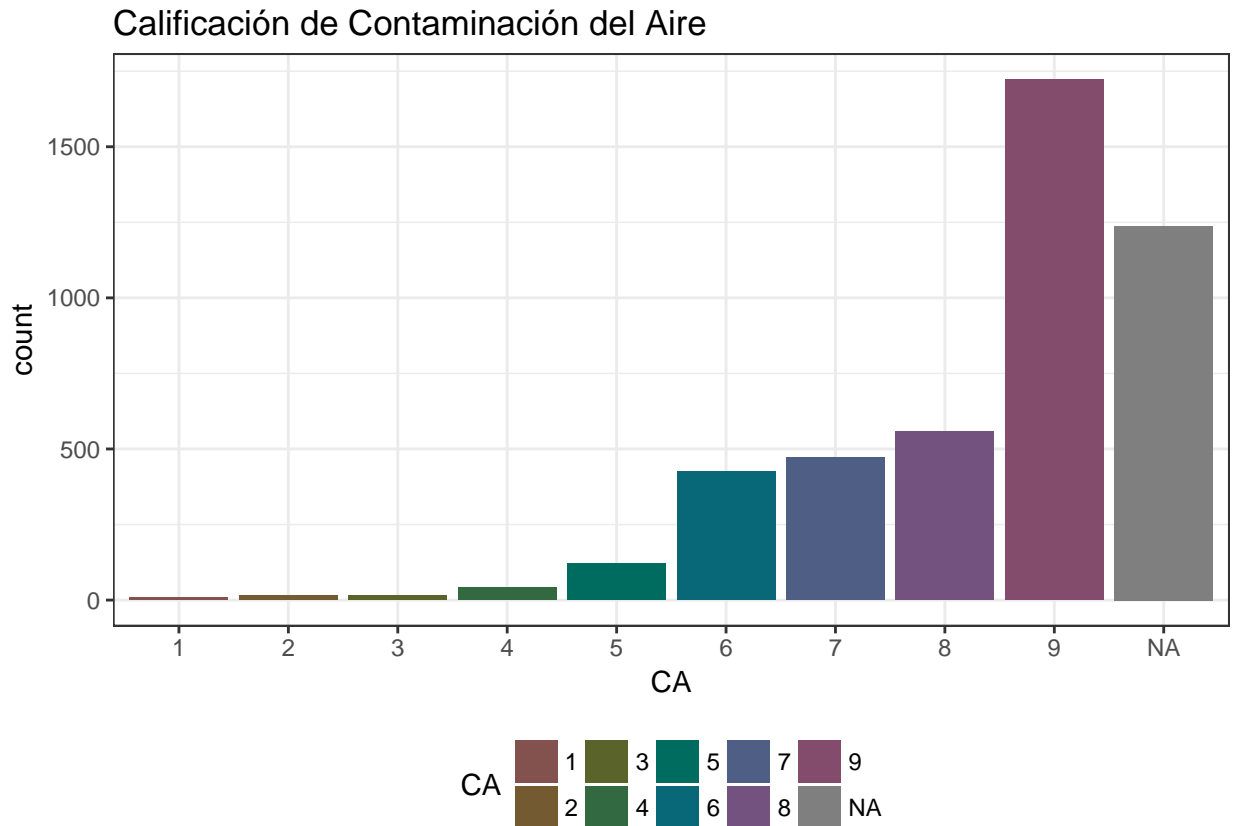
Realizadas las correcciones y ajustes que se han estimado adecuadas para la preparación de los datos, antes de proceder con su procesamiento, se sigue la exportación de una versión depurada.

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

### 4.4 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 contaminantes.

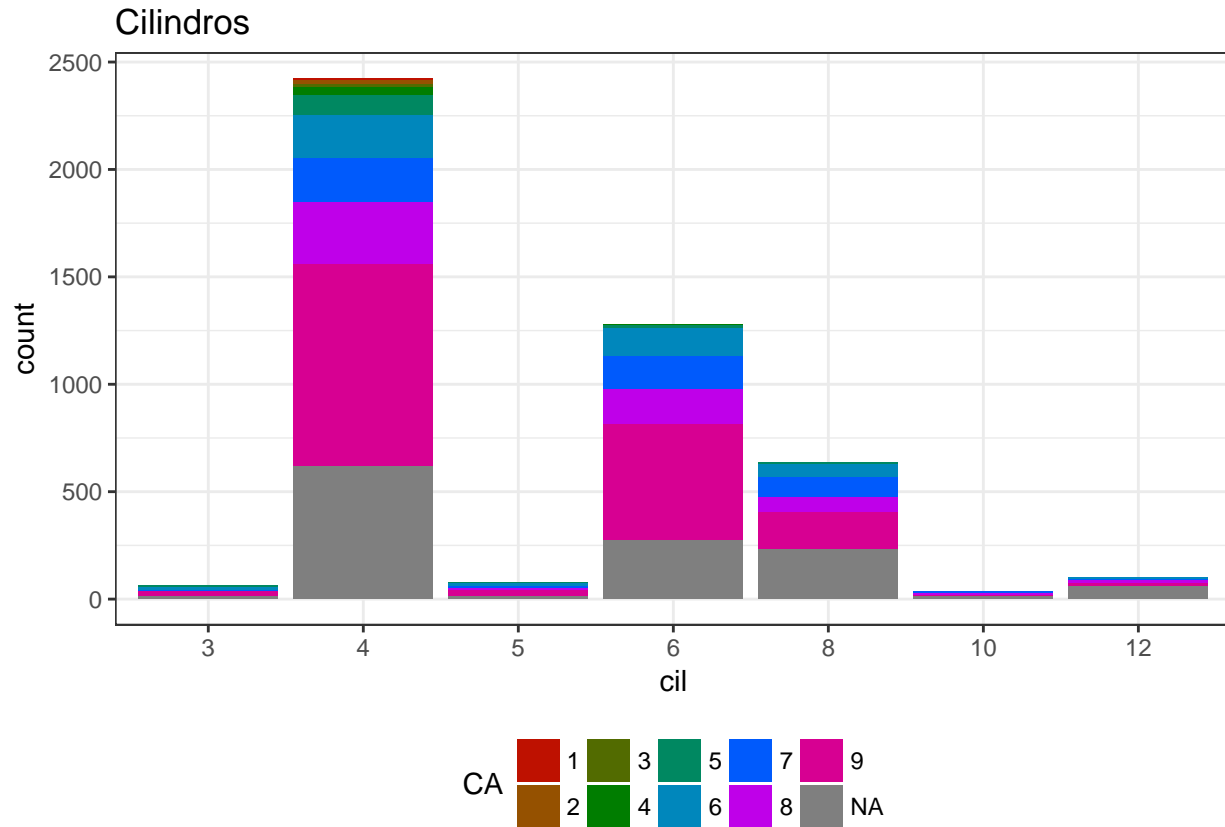
## 4.5 Distribución de Variables Categóricas

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

###

```
ggplot(data = cga2018, aes(x = cil, y = ..count.., fill = CA)) +
  geom_bar() +
  scale_fill_hue(l=40, c=135) +
  labs(title = "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  4  5  6  7  8  9
##          3  0  0  0  0  6 10  5  4 22
##          4  8 16 14 39 93 197 205 292 938
##          5  0  0  0  0  1 17  9  7 32
##          6  0  0  0  2 15 132 153 161 541
```



```
##          8      0      0      0      0      7      62      91      72      169
##          10     0      0      0      0      0      0      7      9      3
##          12     0      0      0      0      0      8      1      14      17
```

```
# Tabulación de proporciones
```

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

```
##
##          1      2      3      4      5      6      7      8      9
##    3  0.00  0.00  0.00  0.00  0.13  0.21  0.11  0.09  0.47
##    4  0.00  0.01  0.01  0.02  0.05  0.11  0.11  0.16  0.52
##    5  0.00  0.00  0.00  0.00  0.02  0.26  0.14  0.11  0.48
##    6  0.00  0.00  0.00  0.00  0.01  0.13  0.15  0.16  0.54
##    8  0.00  0.00  0.00  0.00  0.02  0.15  0.23  0.18  0.42
##   10  0.00  0.00  0.00  0.00  0.00  0.00  0.37  0.47  0.16
##   12  0.00  0.00  0.00  0.00  0.00  0.20  0.02  0.35  0.42
```

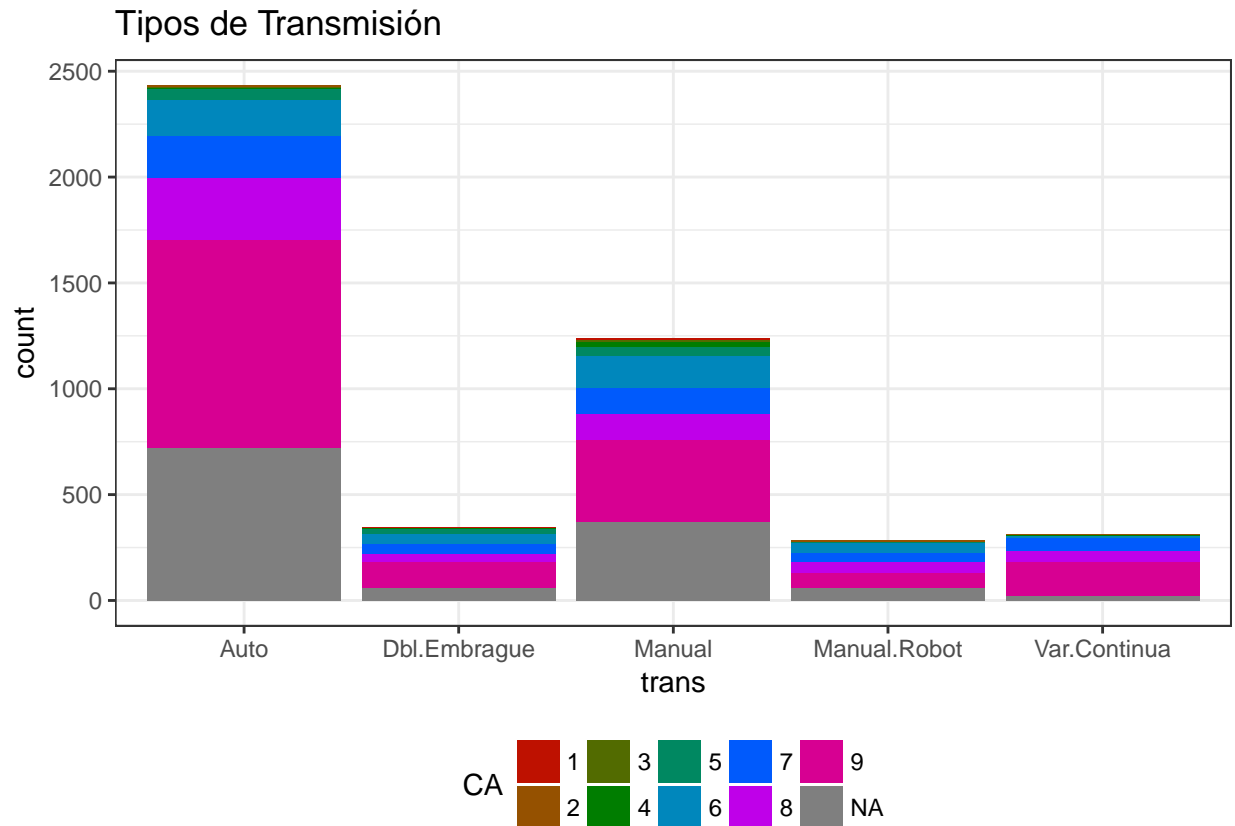
Para el caso de la variable relacionada al *Número de Cilindros* se hace evidente el desequilibrio de la distribución por categoría de *Contaminación de Aire* (claramente sesgado hacia los valores altos de la clasificación).

La gráfica por Tipo de Transmisión, presenta un comportamiento similar a la anterior, con una concentración mayor de modelos contruidos con cajas de cambio Automáticas; pero con una distribución dispersa 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  4  5  6  7  8  9
##      Auto           0  7  5  4  53 168 199 295 981
##      Dbl.Embrague   1  0  0  9  21  50  45  38 124
##      Manual         7  0  8  25  40 153 121 124 388
##      Manual.Robot   0  6  0  0  7  47  40  52  72
##      Var.Continua   0  3  1  3  1  8  66  50 157
```

*# Tabulación de proporciones*

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

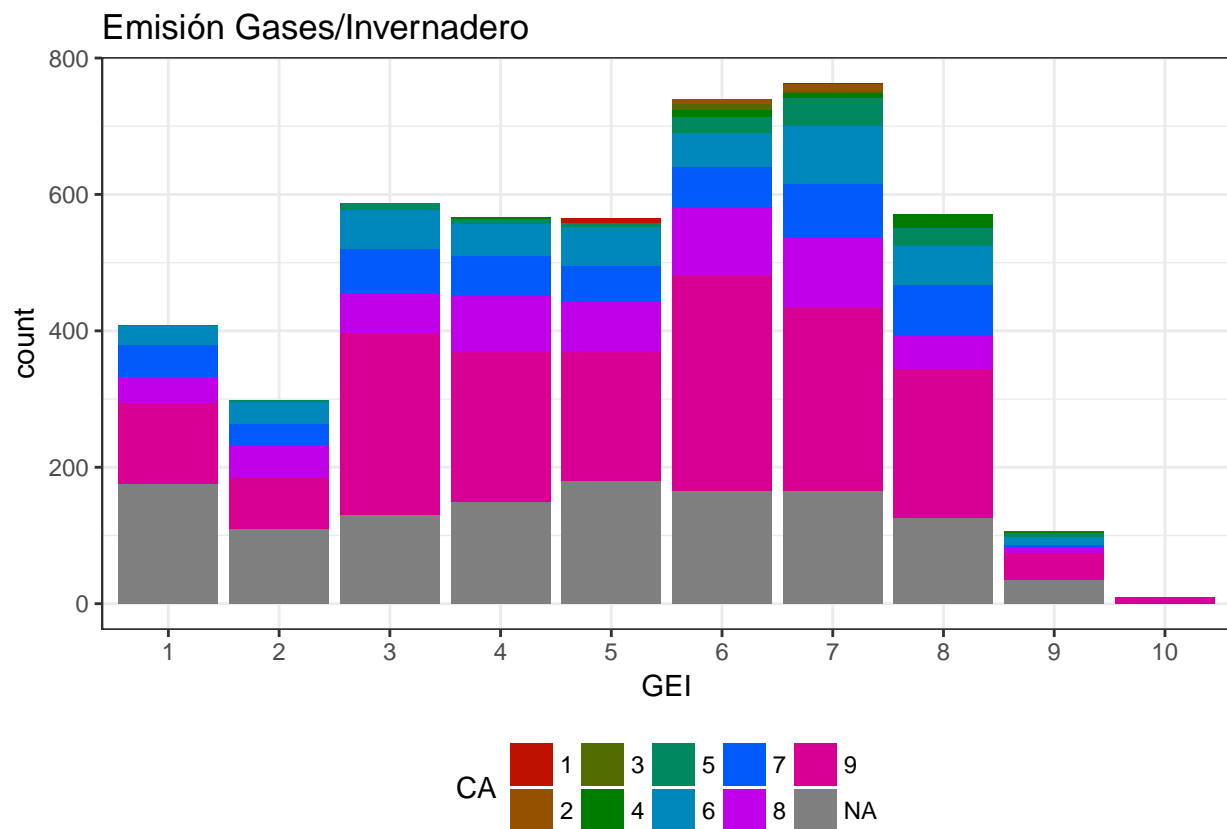
```
##
##      1  2  3  4  5  6  7  8  9
## Auto  0.00 0.00 0.00 0.00 0.03 0.10 0.12 0.17 0.57
## Dbl.Embrague 0.00 0.00 0.00 0.03 0.07 0.17 0.16 0.13 0.43
## Manual  0.01 0.00 0.01 0.03 0.05 0.18 0.14 0.14 0.45
## Manual.Robot 0.00 0.03 0.00 0.00 0.03 0.21 0.18 0.23 0.32
## Var.Continua 0.00 0.01 0.00 0.01 0.00 0.03 0.23 0.17 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 menos intenso.

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 las tabulaciones de frecuencia y proporciones:

### Frecuencia y Proporciones de la distribución del atributo CA sobre la variable GEI  
###

```
# Tabla de frecuencias relativas a la Contaminación del Aire por categoría de Emisión de GEI
table(cga2018$GEI,
      cga2018$CA,
      dnn=c("Emisión de Gases EI", "Contaminación del Aire"))
```

```
##
## Emisión de Gases EI      Contaminación del Aire
##      1      2      3      4      5      6      7      8      9
```

```
##           1      0      0      0      0      1      27      48      37      119
##           2      0      0      0      0      2      32      32      47      76
##           3      0      0      0      0     10      58      64      59     266
##           4      0      0      0      2      6      47      60      82     219
##           5      7      0      0      0      6      58      50      75     190
##           6      0      7      9     10     24      49      60     100     315
##           7      1      9      4      7     41      85      80     102     268
##           8      0      0      0     20     26      58      74      49     218
##           9      0      0      1      2      6     12       3       8      41
##          10      0      0      0      0      0      0       0       0      10
```

```
# Tabulación de proporciones
```

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

```
##
##           1      2      3      4      5      6      7      8      9
##    1  0.00 0.00 0.00 0.00 0.00 0.12 0.21 0.16 0.51
##    2  0.00 0.00 0.00 0.00 0.01 0.17 0.17 0.25 0.40
##    3  0.00 0.00 0.00 0.00 0.02 0.13 0.14 0.13 0.58
##    4  0.00 0.00 0.00 0.00 0.01 0.11 0.14 0.20 0.53
##    5  0.02 0.00 0.00 0.00 0.02 0.15 0.13 0.19 0.49
##    6  0.00 0.01 0.02 0.02 0.04 0.09 0.10 0.17 0.55
##    7  0.00 0.02 0.01 0.01 0.07 0.14 0.13 0.17 0.45
##    8  0.00 0.00 0.00 0.04 0.06 0.13 0.17 0.11 0.49
##    9  0.00 0.00 0.01 0.03 0.08 0.16 0.04 0.11 0.56
##   10  0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
```

## 4.6 Distribución de Variables Continuas

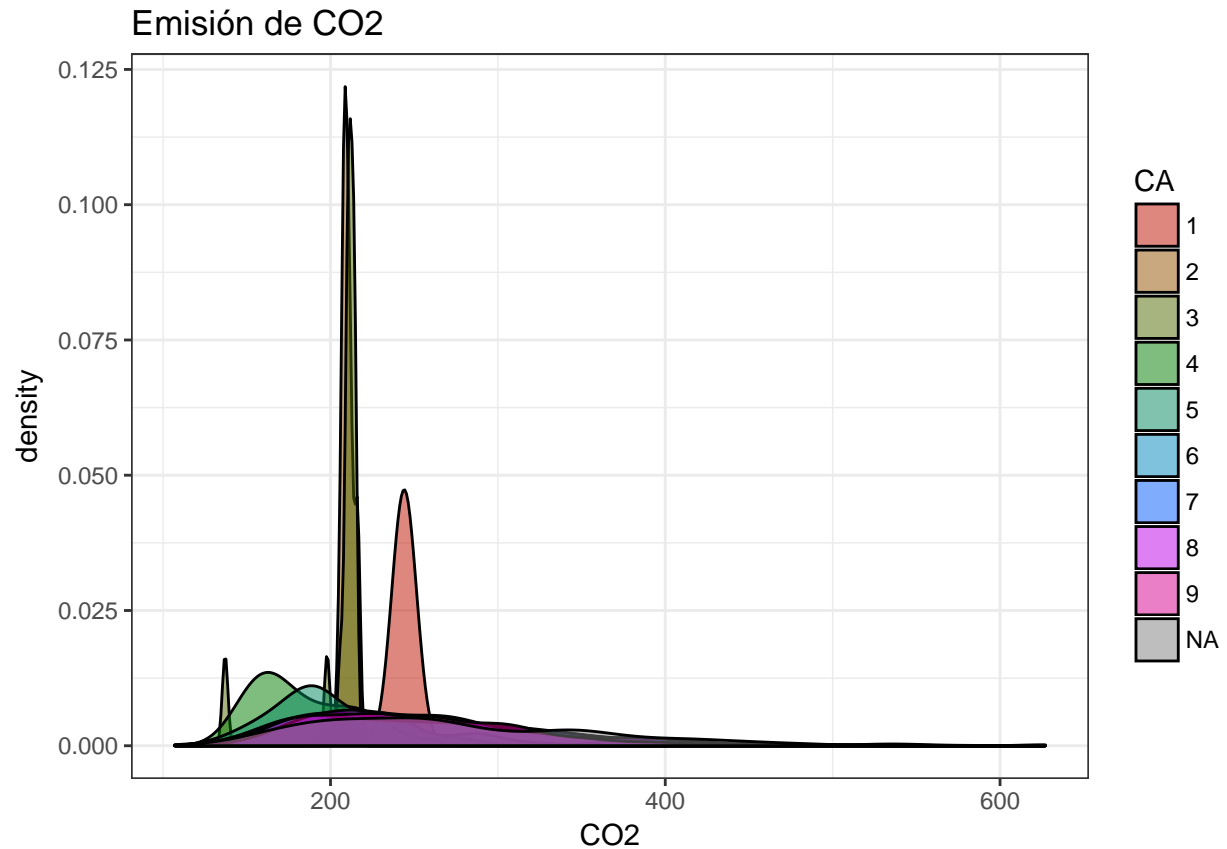
Se presume la existencia de una relación (de mucha relevancia) entre la variable de respuesta para la calificación de la (\_\_\*\* Contaminación del Aire), *y los atributos continuos relacionados a la medición de emisiones a la atmosfera (CO2\*\*\_\_ y NOx).*

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()
```

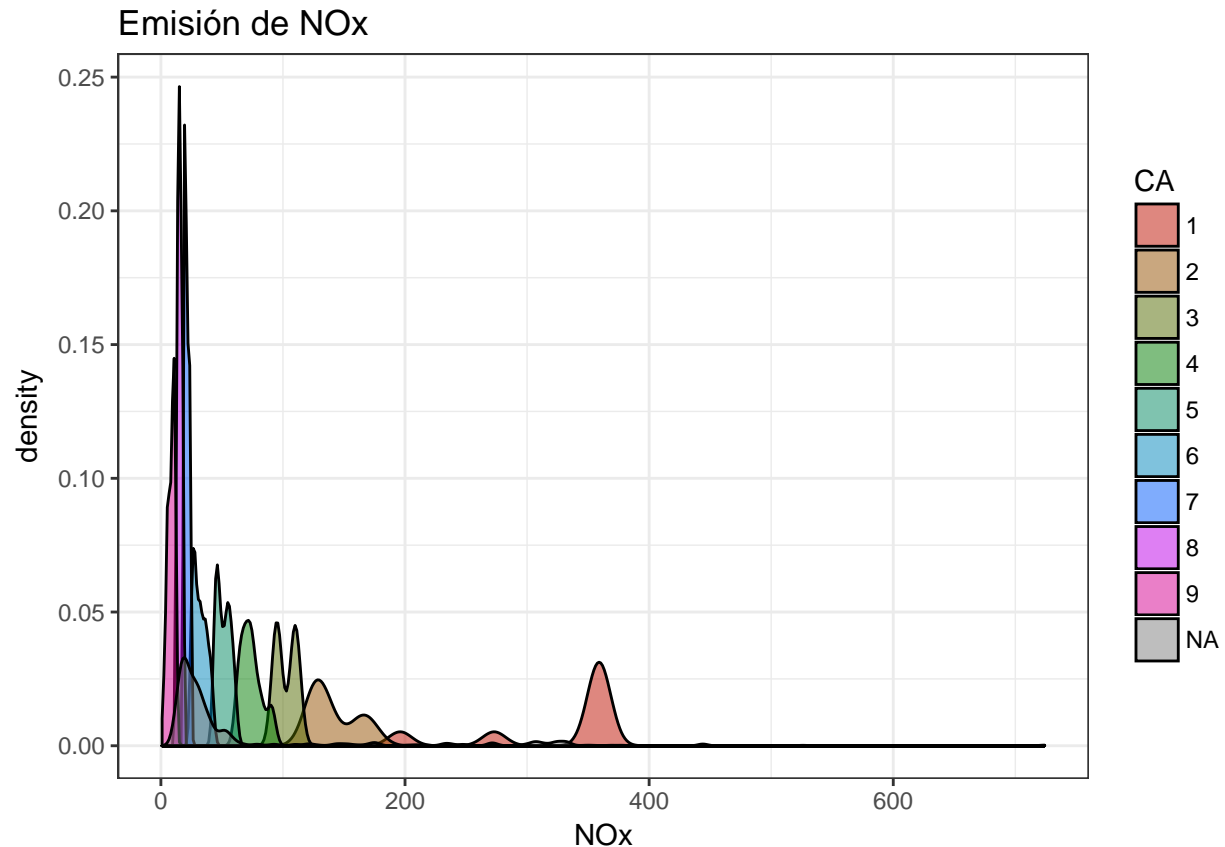


Se aprecian formas de distribuciones bastante dispares entre categorías de la variable de respuesta, quizá producto de la pronunciada diferencia en la cantidad de casos por clase; pero con una tendencia a concentrarse en un entorno del valor de los 200 gr/Km.

### Gráfica de densidad de NOx sobre categorías 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).

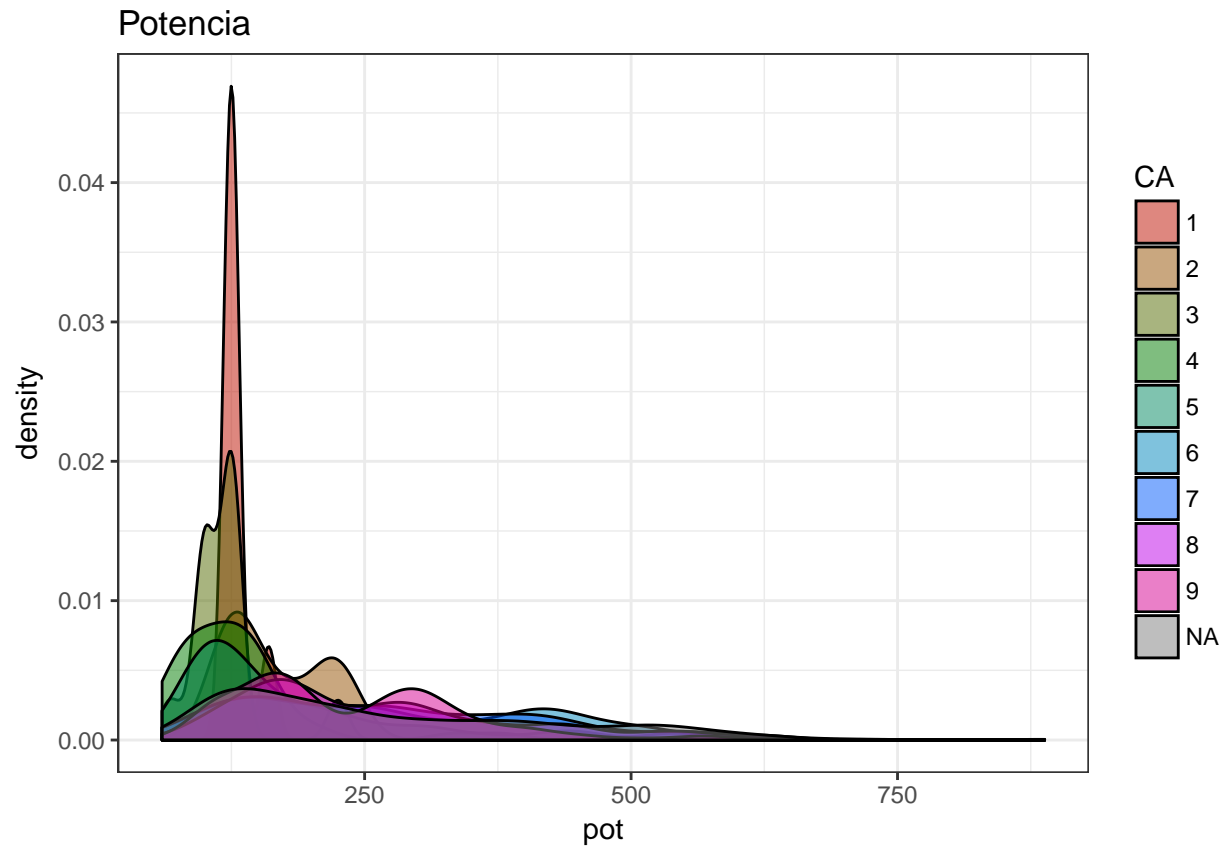
Extendiendo la inspección gráfica al resto de las variables continuas de interés, tenemos:

La variable **potencia** exhibe un comportamiento similar, con formas de distribución dispares 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()
```

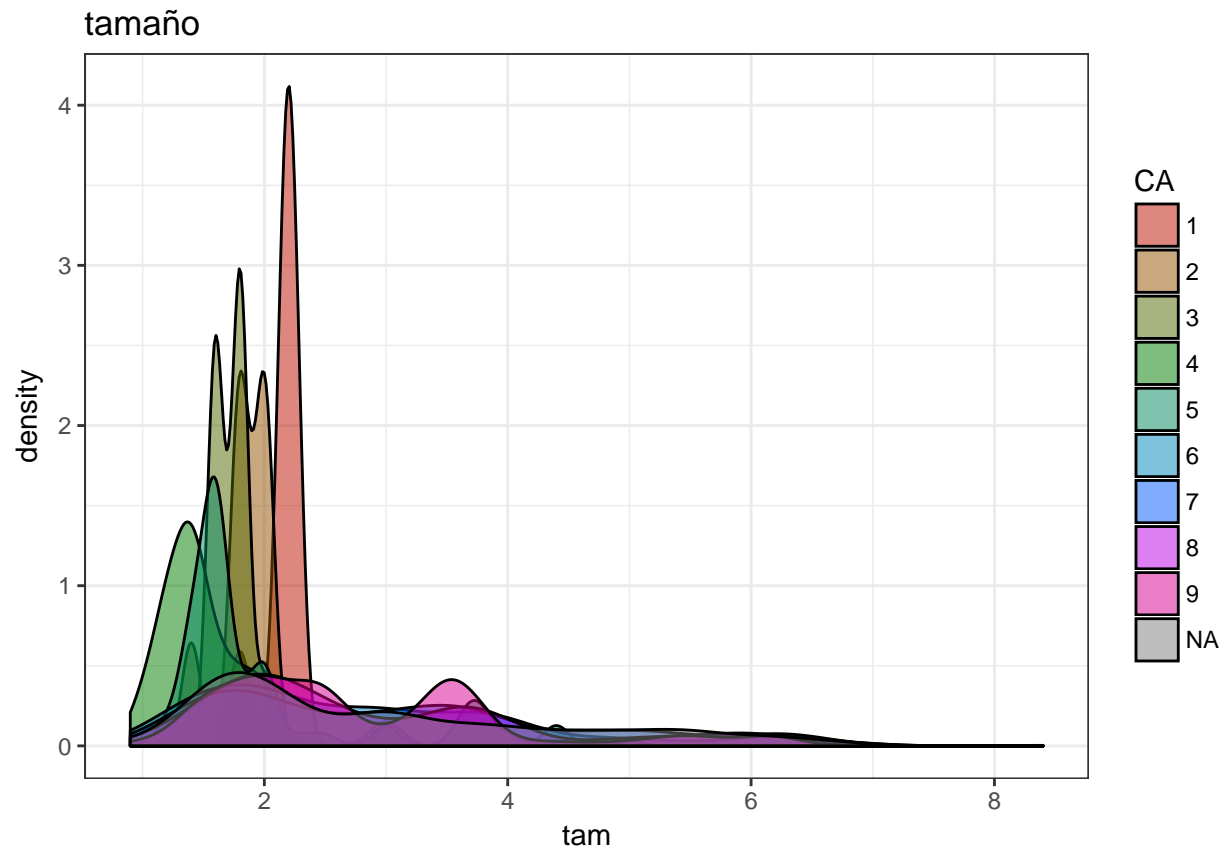


El atributo *tam* no es la excepción, y muestra un conducta similar a las anteriores, mostrando que la oferta de modelos está concentrada en motores con capacidad de alrededor de 2 litros.

### Gráfica de densidad de tam sobre categorías del atributo CA

###

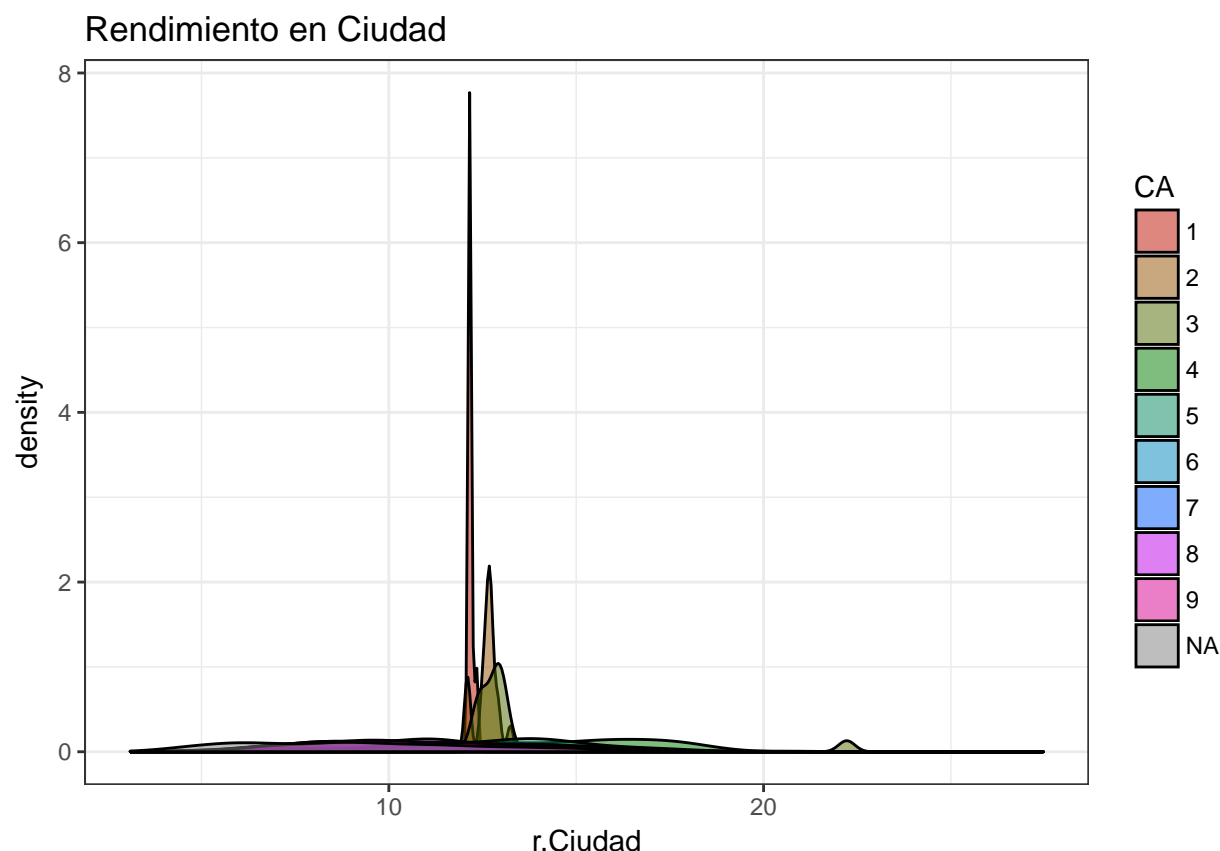
```
ggplot(data = cga2018, aes(x = tam, fill = CA)) +  
  geom_density(alpha = 0.5) +  
  scale_fill_hue(l=40, c=135) +  
  labs(title = "tamaño") +  
  theme_bw()
```



Finalmente, para el *rendimient en Ciudad* tenemos una situación muy paracida, con bastante más difererncia 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()
```





## 4.7 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 la muy característica ausencia de valores (inicializados a cero) para un segmento importante de 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	0	0	0	0	7	0	1	0	0
##	2	0	0	0	0	0	7	9	0	0
##	3	0	0	0	0	0	9	4	0	1
##	4	0	0	0	2	0	10	7	20	2
##	5	1	2	10	6	6	24	41	26	6
##	6	27	32	58	47	58	49	85	58	12
##	7	48	32	64	60	50	60	80	74	3
##	8	37	47	59	82	75	100	102	49	8
##	9	119	76	266	219	190	315	268	218	41
##	10	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 serán 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 de niveles 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
##      Db1.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
##      Db1.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**.

Colateralmente, esta medida contribuye 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 ...
## $ cil        : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<...: 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/ 9 levels "1","2","3","4",...: 9 9 9 9 9 9 9 9 9 ...
```

## 4.8 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 los resultados.

```
### 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 los grupos de datos a analizar/comparar

Habiendo excluido el *Tipo de Combustible*, las variables categóricas restantes *Tipo de Transmisión*, *Número de Cilindros* y *Emisión de Gases de Efecto Invernadero*, pueden ser utilizadas para la

conformación de grupos que puedan ayudar a entender el comportamiento del modelo.

*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 resulten de la segmentación por variable categórica.
- 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 la normalidad y homogeneidad de la 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 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

Considerando que los grupos a inspeccionar, determinados por las variables categóricas *Tipo de Transmisión*, *Número de Cilindros* y *Emisión de Gases de Efecto Invernadero*, se aplicará la prueba de Levene de Homogeneidad de Varianza, sobre las variables continuas *pot* y *NOx*, que son las que hasta ahora parecen prometedoras en relación a su capacidad predictiva en relación a la calificación de *Contaminación de Aire*

```
### Evaluación de Levene para la Potencia por atributo Categórico
###
leveneTest(y=cga2018.dat$pot, group=cga2018.dat$trans, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value    Pr(>F)
## group  4  38.773 < 2.2e-16 ***
##      3366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y=cga2018.dat$pot, group=cga2018.dat$cil, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value    Pr(>F)
## group  6  27.92 < 2.2e-16 ***
##      3364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y=cga2018.dat$pot, group=cga2018.dat$GEI, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value    Pr(>F)
## group  9  63.277 < 2.2e-16 ***
##      3361
## ---
## 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*, para los diferentes establecidos, resultan en valores para el 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 continua para cada uno de los grupos.

```
### Evaluación de Levene para NOx por atributo Categórico
###
leveneTest(y=cga2018.dat$NOx, group=cga2018.dat$trans, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value    Pr(>F)
## group  4  14.204 1.686e-11 ***
##      3366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## Levene's Test for Homogeneity of Variance (center = "median")
```

```
##           Df F value    Pr(>F)
## group      6 11.612 6.553e-13 ***
##           3364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
## Levene's Test for Homogeneity of Variance (center = "median")
##           Df F value    Pr(>F)
## group      9 11.565 < 2.2e-16 ***
##           3361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para el caso de la variable *NOx*, observamos el mismo resultado.

Finalmente, para las pruebas sobre el *rendimiento en Ciudad* se tiene:

```
### Evaluación de Levene para r.Ciudad por atributo Categórico
###
leveneTest(y=cga2018.dat$r.Ciudad, group=cga2018.dat$trans, center="median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##           Df F value    Pr(>F)
## group      4 7.8763 2.575e-06 ***
##           3366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## Levene's Test for Homogeneity of Variance (center = "median")
##           Df F value    Pr(>F)
## group      6 46.088 < 2.2e-16 ***
##           3364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## Levene's Test for Homogeneity of Variance (center = "median")
##           Df F value    Pr(>F)
## group      9 28.669 < 2.2e-16 ***
##           3361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En conclusión, para todos los casos estudiados, la variable continua tiene varianza distinta en al menos alguno de los niveles, en el dominio de cada variable categórica utilizada para la definición de los grupos de interés.

Por lo tanto, en todos los casos probados, se rechaza la hipótesis nula de la prueba, en beneficio de la alternativa que establece la heteroscedasticidad de las variables para cada uno de los grupos.

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

```
##           transmission      cilindros      GEI
## field.pot 6.679976e-38 2.945863e-26 1.725949e-102
## field.tam 1.076318e-78 1.462485e-127 6.430449e-113
```

```
## field.cdad 3.544466e-08 5.689959e-63 2.088101e-42
## field.co2 1.220280e-40 2.576000e-49 9.834100e-186
## field.nox 3.478426e-26 1.153527e-09 4.914277e-22
```

### 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 la anova, para validar la homogeneidad de las medias a través de los grupos de cada 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 componentes; y 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 rho, por lo que se ensaya el método de Kendall como alternativa, explorando la determinación de tau-b 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=9.338839e-01 (tau=1.078775e-03) (r^2=1.163756e-06) *** "
## [1] "      tam: p-vle=1.599677e-06 (tau=6.401114e-02) (r^2=4.097426e-03)      "
## [1] "    r.Ciudad: p-vle=3.367498e-06 (tau=-6.029133e-02) (r^2=3.635045e-03)  "
## [1] "r.Carretera: p-vle=1.530222e-05 (tau=-5.608931e-02) (r^2=3.146011e-03)  "
## [1] "      r.Comb: p-vle=3.219285e-03 (tau=-3.817788e-02) (r^2=1.457550e-03)  "
## [1] "    r.Ajust: p-vle=3.183618e-03 (tau=-3.822692e-02) (r^2=1.461298e-03)  "
## [1] "      CO2: p-vle=3.270294e-03 (tau=3.818403e-02) (r^2=1.458020e-03)      "
## [1] "      NOx: p-vle=0.000000e+00 (tau=-8.373843e-01) (r^2=7.012124e-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 de 0.7 para el coeficiente de correlación, 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.

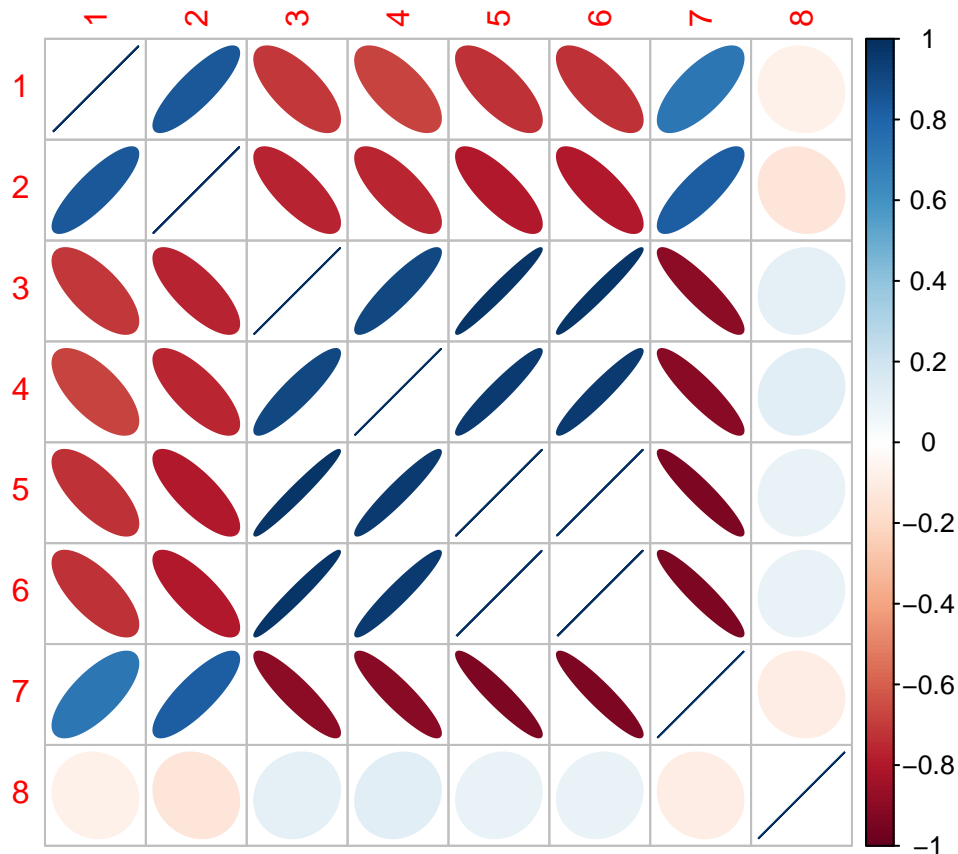
### 5.3.2 Correlacionadas entre variables continuas

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 <- cbind(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)

M <- cor(aux)
corrplot(M, method = "ellipse")
```





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, 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 relación de 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 correspondiente a la salida se lista de seguido:

```
### Inspección de la Media por Categoría de Respuesta - atr C02
###
for (i in 1:length(levels(cga2018.dat$CA))) {
  j <- levels(cga2018.dat$CA)[i]
  print( paste( "Contaminación Aire: cat. ",
               sprintf("%s %6.2f", j, mean(cga2018.dat$C02[cga2018.dat$CA == j]) ) ) )
}
```

```
## [1] "Contaminación Aire: cat. 1 102.00"
## [1] "Contaminación Aire: cat. 2 103.44"
## [1] "Contaminación Aire: cat. 3 103.62"
## [1] "Contaminación Aire: cat. 4 79.51"
```

```
## [1] "Contaminación Aire: cat. 5 110.61"
## [1] "Contaminación Aire: cat. 6 147.01"
## [1] "Contaminación Aire: cat. 7 152.15"
## [1] "Contaminación Aire: cat. 8 149.71"
## [1] "Contaminación Aire: cat. 9 143.83"
```

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: cat. ",
              sprintf("%s %6.2f", j, mean(cga2018.dat$NOx[cga2018.dat$CA == j]) ) ) )
}
```

```
## [1] "Contaminación Aire: cat. 1 195.00"
## [1] "Contaminación Aire: cat. 2 140.44"
## [1] "Contaminación Aire: cat. 3 101.15"
## [1] "Contaminación Aire: cat. 4 73.00"
## [1] "Contaminación Aire: cat. 5 50.35"
## [1] "Contaminación Aire: cat. 6 31.18"
## [1] "Contaminación Aire: cat. 7 20.23"
## [1] "Contaminación Aire: cat. 8 14.45"
## [1] "Contaminación Aire: cat. 9 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: cat. ",
              sprintf("%s %6.2f", j, mean(cga2018.dat$r.Ciudad[cga2018.dat$CA == j]) ) ) )
}
```

```
## [1] "Contaminación Aire: cat. 1 8.34"
## [1] "Contaminación Aire: cat. 2 8.61"
## [1] "Contaminación Aire: cat. 3 8.75"
## [1] "Contaminación Aire: cat. 4 10.75"
## [1] "Contaminación Aire: cat. 5 8.69"
## [1] "Contaminación Aire: cat. 6 6.90"
## [1] "Contaminación Aire: cat. 7 6.86"
## [1] "Contaminación Aire: cat. 8 6.71"
## [1] "Contaminación Aire: cat. 9 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 más homogéneos.

Para este último caso es prudente verificar si estadísticamente 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 por debajo de la categoría 5, 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
## y el resto por debajo de 5
##
md.cdad.bajos <- subset(cga2018.dat, CA == 1 |
                        CA == 2 |
                        CA == 3 |
                        CA == 4 |
                        CA == 5 ,
                        select=c(r.Ciudad))

md.cdad.altos <- subset(cga2018.dat, CA == 6 |
                       CA == 7 |
                       CA == 8 |
                       CA == 9 |
                       CA == 10 ,
                       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.

El resultado puede estar influido por la distribución de observaciones por categoría de **CA** que hemos ya identificado en secciones previas.

### *Varían entre si las medias de a pares de categorías, para las categorías superiores?*

Para este tipo de preguntas pudiera aplicarse el análisis de varianza, 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 == 6, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 7, select=c(r.Ciudad))

cat.6.vs.7 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA == 7, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 8, select=c(r.Ciudad))
```

```

cat.7.vs.8 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA == 8, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA == 9, select=c(r.Ciudad))

cat.8.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA == 6, select=c(r.Ciudad))
md.cdad.altos <- subset(cga2018.dat, CA ==9, select=c(r.Ciudad))

cat.6.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

rslt <- cbind(cat.6.vs.7$p.value, cat.7.vs.8$p.value, cat.8.vs.9$p.value, cat.6.vs.9$p.value)
colnames(rslt) <- c("6.vs.7", "7.vs.8", "8.vs.9", "6.vs.9")
rslt

```

```

##          6.vs.7    7.vs.8    8.vs.9    6.vs.9
## [1,] 0.8591474 0.4132481 0.1411686 0.9233426

```

El resultado para cada una de las comparaciones da por encima del nivel de significación, por lo que no puede rechazarse la igualdad de media entre estas categorías.

*Varia la media del CO2 para el grupo de autos con CA por debajo de la categoría 5, con respecto al resto (para el subgrupo de autos de cuatro cilindros) ?*

La pregunta va en línea con la intención original de utilizar las variables categóricas seleccionadas de la muestra para la evaluación de grupos 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
## y el resto por debajo de 5, en coches de 4 cilindros
##
md.cdad.bajos <- subset(cga2018.dat, (CA==1|CA==2|CA==3|CA==4|CA==5) & (cil==4),select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, (CA==6|CA==7|CA==8|CA==9|CA==10) & (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 altas).

*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==6 & cil==4, select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, CA==7 & cil==4, select=c(CO2))

cat.6.vs.7 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==7 & cil==4, select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, CA==8 & cil==4, select=c(CO2))

cat.7.vs.8 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==8 & cil==4, select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, CA==9 & cil==4, select=c(CO2))

cat.8.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==6 & cil==4, select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, CA==9 & cil==4, select=c(CO2))

cat.6.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==7 & cil==4, select=c(CO2))
md.cdad.altos <- subset(cga2018.dat, CA==9 & cil==4, select=c(CO2))

cat.7.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

rslt <- cbind(cat.6.vs.7$p.value,
              cat.7.vs.8$p.value,
              cat.8.vs.9$p.value,
              cat.6.vs.9$p.value,
              cat.7.vs.9$p.value)

colnames(rslt) <- c("6.vs.7", "7.vs.8", "8.vs.9", "6.vs.9", "7.vs.9")
rslt
```

```
##           6.vs.7           7.vs.8           8.vs.9           6.vs.9           7.vs.9
## [1,] 0.08904701 5.665669e-05 0.05681967 0.5342004 0.00238586
```

A excepción de la categoría 7, que durante su comparación con las 8 y 9, obtiene p-valores que permiten afirmar que tienen medias distintas, para el resto de las comparaciones no hay evidencia de las medias sean significativamente distintas para las categorías 6, 8 y 9.

*Dentro del grupo de categorías altas para CA serán distintas las medias del atributo NOx entre las categorías, para coches con transmisión automática ?*

El poco poder predictivo que intuimos 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:

```
#### Comparación a pares de la media del atributo NOx para los grupos superiores
##   en autos de transmisión automática
```

```
##
md.cdad.bajos <- subset(cga2018.dat, CA==6 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==7 & trans=="Auto", select=c(NOx))

cat.6.vs.7 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==7 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==8 & trans=="Auto", select=c(NOx))

cat.7.vs.8 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==8 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==9 & trans=="Auto", select=c(NOx))

cat.8.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==6 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==9 & trans=="Auto", select=c(NOx))

cat.6.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

md.cdad.bajos <- subset(cga2018.dat, CA==7 & trans=="Auto", select=c(NOx))
md.cdad.altos <- subset(cga2018.dat, CA==9 & trans=="Auto", select=c(NOx))

cat.7.vs.9 <- t.test( md.cdad.bajos, md.cdad.altos, var.equal=FALSE, paired=FALSE)

rslt <- cbind(cat.6.vs.7$p.value,
              cat.7.vs.8$p.value,
              cat.8.vs.9$p.value,
              cat.6.vs.9$p.value,
              cat.7.vs.9$p.value)

colnames(rslt) <- c("6.vs.7", "7.vs.8", "8.vs.9", "6.vs.9", "7.vs.9")
rslt
```

```
##           6.vs.7           7.vs.8           8.vs.9           6.vs.9           7.vs.9
## [1,] 5.343364e-65 5.161067e-124 5.843811e-246 1.349579e-121 4.353678e-258
```

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.

## 6. Modelo Tentativo

En el siguiente segmento se procede a extraer una parte del conjunto de datos para utilizarlo como set de pruebas, para validar la precisión del modelo generado.

```
#### 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

### Muestreo aleatorio para extraer el set de Entrenamiento
###
```

```

set.seed(1965)

m.training.set <- sample( sample.size, sample.size * m.training.size )

m.training <- cga2018.dat[ m.training.set, ]
m.test <- cga2018.dat[ -m.training.set, ]

# Homologación de los niveles de test
# por la baja representatividad de la categoria en los datos
# debemos asegurarnos que el niveleste en el caonjunto de test
# a modo que las predicciones puedan ser comparable con el modelo

levels(m.test$CA) <- c(levels(m.test$CA), "10")

m.test.rsl <- m.test$CA

## se eliminan los atributos redundantes o que han demostrado baja contribución
## por su escaso poder predictivo, además de la variable objetivo
##

m.training_1 <- m.training[ , c( -3, -4, -5, -6, -7, -8, -12)]
m.test_1 <- m.test[ , c( -3, -4, -5, -6, -7, -8, -12)]

str(m.training_1)

## 'data.frame': 2696 obs. of 5 variables:
## $ trans: Factor w/ 5 levels "Auto","Dbl.Embrague",...: 1 1 3 3 1 4 1 3 1 5 ...
## $ cil : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<...: 2 7 4 2 3 2 2 4 5 2 ...
## $ CO2 : num 142 518 246 90 121 68 159 321 424 101 ...
## $ NOx : num 14 4 13 30 3 15 13 40 13 13 ...
## $ GEI : Factor w/ 10 levels "1","2","3","4",...: 5 1 1 7 6 8 4 1 3 7 ...

```

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 kmeans 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 existen 10 categorías registradas para el atributo **CA**, impondremos  $k = 10$  para la inicialización del hiperparámetro del modelo.

```

#### kmeans para data mixta: numérica y categórica
##
mv <- kproto(m.training_1, 10)

## # 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_1)
```

```
m.predicted.clusters <- predict(mv, m.test_1)
m.pred <- factor(m.predicted.clusters$cluster)
```

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

```
## Confusion Matrix and Statistics
```

```
##
```

```
##           Reference
```

```
## Prediction  1  2  3  4  5  6  7  8  9 10
```

```
##           1  0  1  1  0  0  0  0  0  0
```

```
##           2  0  0  0  8  4  9 10 13 54
```

```
##           3  0  0  0  0  0  0  3  7 36
```

```
##           4  0  0  0  0  0 10 11 10 42
```

```
##           5  0  0  0  0  0  1  7  7 34
```

```
##           6  0  0  0  1 11 19 15  9 25
```

```
##           7  0  0  0  0  0  2 11 18 44
```

```
##           8  0  0  2  1  4  4  0  0  0
```

```
##           9  0  0  0  0  0 10 10 10 29
```

```
##          10  0  0  0  0  1 18 33 39 91
```

```
##
```

```
## Overall Statistics
```

```
##
```

```
##           Accuracy : 0.0874
```

```
##           95% CI : (0.0672, 0.1113)
```

```
## No Information Rate : 0.5259
```

```
## P-Value [Acc > NIR] : 1
```

```
##
```

```
##           Kappa : 0.0056
```

```
## McNemar's Test P-Value : NA
```

```
##
```

```
## Statistics by Class:
```

```
##
```

```
##           Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
```

```
## Sensitivity           NA 0.000000 0.000000 0.000000 0.000000 0.26027
```

```
## Specificity           0.997037 0.854599 0.931548 0.89023 0.92519 0.89867
```

```
## Pos Pred Value         NA 0.000000 0.000000 0.000000 0.000000 0.23750
```

```
## Neg Pred Value         NA 0.998267 0.995231 0.98339 0.96805 0.90924
```

```
## Prevalence             0.000000 0.001481 0.004444 0.01481 0.02963 0.10815
```

```
## Detection Rate         0.000000 0.000000 0.000000 0.000000 0.000000 0.02815
```

```
## Detection Prevalence 0.002963 0.145185 0.068148 0.10815 0.07259 0.11852
```

```
## Balanced Accuracy      NA 0.427300 0.465774 0.44511 0.46260 0.57947
```

```
##           Class: 7 Class: 8 Class: 9 Class: 10
```

```
## Sensitivity           0.1100 0.0000 0.08169 NA
```

```
## Specificity           0.8887 0.9804 0.90625 0.7304
```

```
## Pos Pred Value         0.1467 0.0000 0.49153 NA
```

```
## Neg Pred Value         0.8517 0.8298 0.47078 NA
```

```
## Prevalence             0.1481 0.1674 0.52593 0.0000
```

```
## Detection Rate         0.0163 0.0000 0.04296 0.0000
```

```
## Detection Prevalence 0.1111 0.0163 0.08741 0.2696
```

```
## Balanced Accuracy      0.4993 0.4902 0.49397 NA
```

La precisión obtenida es demasiado baja como para promover el modelo, lo que nos reafirma el carácter poco



predictivo de los datos y sesgo que estructuralmente se ha venido apreciando en los datos disponibles a lo largo de las pruebas estadísticas previas.

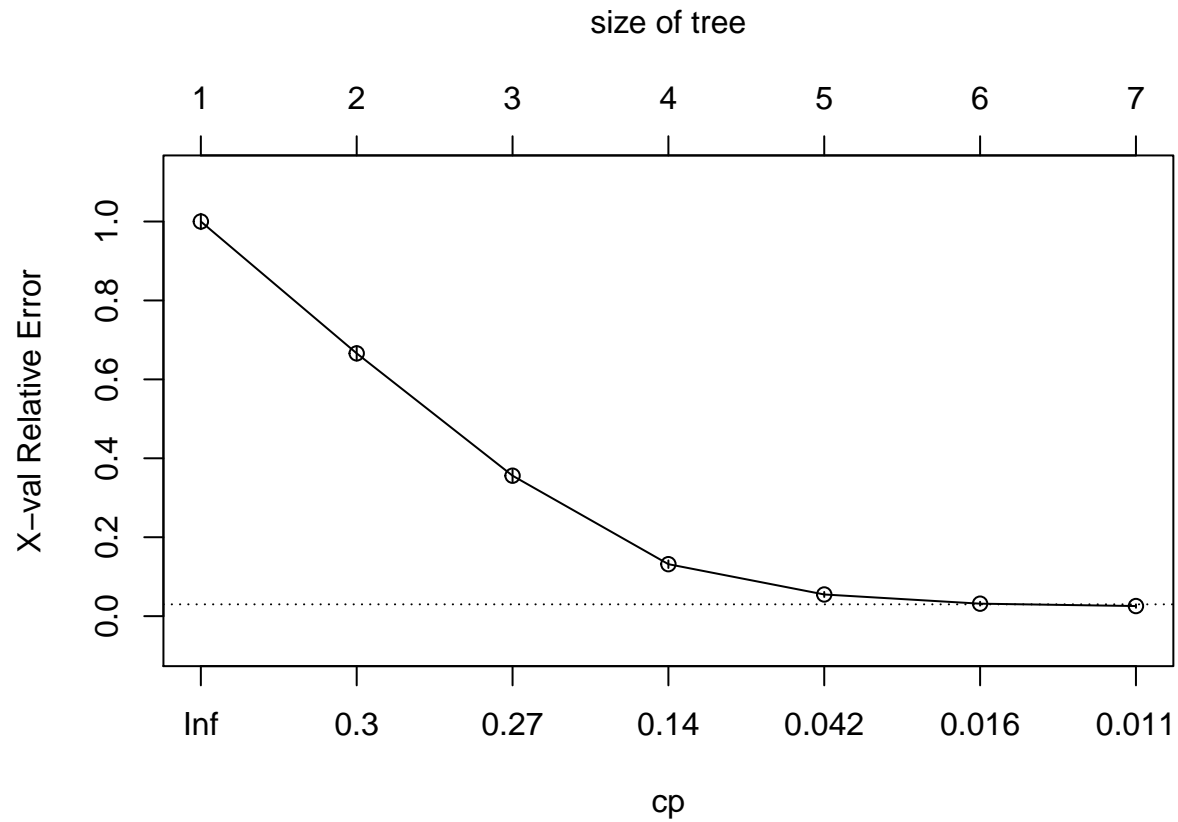
Alternativamente, consideraremos un modelo de árbol que ayude a ilustrar mejor la realidad de las características de nuestro set de datos:

```
#### Modelo de árbol CART
##
mv <- rpart(CA~NOx+CO2+trans+cil+GEI, data=m.training, method="class")

printcp(mv) # resultados del modelo

##
## Classification tree:
## rpart(formula = CA ~ NOx + CO2 + trans + cil + GEI, data = m.training,
##       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 0.334086      0 1.000000 1.000000 0.0195327
## 2 0.271633      1 0.665914 0.665914 0.0183462
## 3 0.262603      2 0.394281 0.355907 0.0148599
## 4 0.076749      3 0.131678 0.131678 0.0096254
## 5 0.023326      4 0.054929 0.054929 0.0063413
## 6 0.011287      5 0.031603 0.031603 0.0048383
## 7 0.010000      6 0.020316 0.025583 0.0043597

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

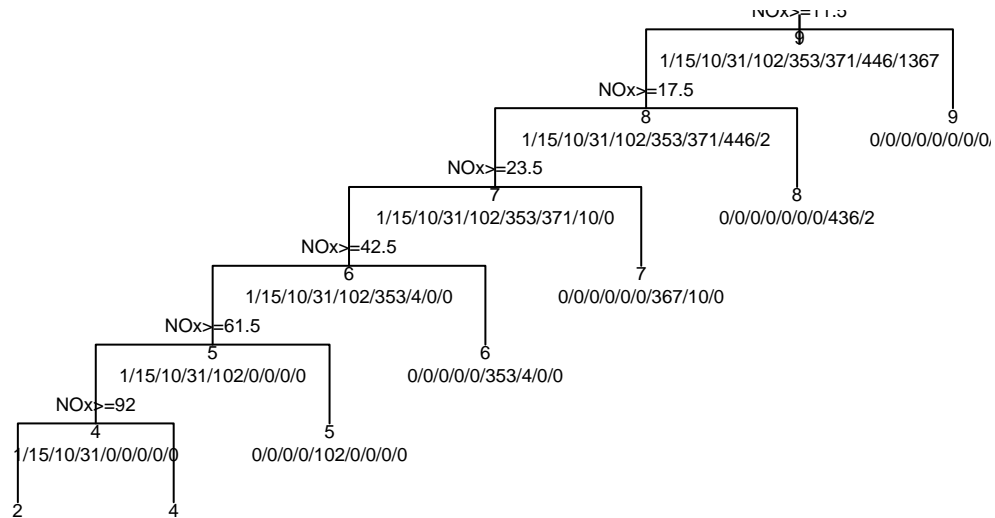


El atributo ***NOx*** es considerado como suficiente para la confección del árbol, cuestión que intuitivamente se anticipaba a partir del estudio de las medias, realizado en secciones previas.

El nivel de complejidad que minimiza el error es que corresponde al CP 7, tal como se ve en la tabulación de resultados y la gráfica correspondiente.

```
#### Modelo de árbol CART
##
plot(mv, uniform=TRUE, main="Clasificación de CA")
text(mv, use.n=TRUE, all=TRUE, cex=.6)
```

## Clasificación de CA



Utilizando el conjunto reservado para el test, y estimar 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_1,
                           type = "class")
```

```
m.pred <- factor(m.rpart.predict)
```

```
#
```

```
# Matriz 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  4  5  6  7  8  9 10
##           1  0  0  0  0  0  0  0  0  0
##           2  0  1  3  0  0  0  0  0  0
##           3  0  0  0  0  0  0  0  0  0
##           4  0  0  0 10  0  0  0  0  0
##           5  0  0  0  0 20  0  0  0  0
```

```

##          6      0      0      0      0      0      73      1      0      0      0
##          7      0      0      0      0      0      0      99      4      0      0
##          8      0      0      0      0      0      0      0     109      0      0
##          9      0      0      0      0      0      0      0      0     355      0
##         10      0      0      0      0      0      0      0      0      0      0
##
## Overall Statistics
##
##              Accuracy : 0.9881
##              95% CI : (0.9768, 0.9949)
##      No Information Rate : 0.5259
##      P-Value [Acc > NIR] : < 2.2e-16
##
##              Kappa : 0.9821
##  McNemar's Test P-Value : NA
##
## Statistics by Class:
##
##              Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
## Sensitivity          NA 1.000000 0.000000 1.00000 1.00000 1.0000
## Specificity          1 0.995549 1.000000 1.00000 1.00000 0.9983
## Pos Pred Value       NA 0.250000      NaN 1.00000 1.00000 0.9865
## Neg Pred Value       NA 1.000000 0.995556 1.00000 1.00000 1.0000
## Prevalence           0 0.001481 0.004444 0.01481 0.02963 0.1081
## Detection Rate       0 0.001481 0.000000 0.01481 0.02963 0.1081
## Detection Prevalence 0 0.005926 0.000000 0.01481 0.02963 0.1096
## Balanced Accuracy     NA 0.997774 0.500000 1.00000 1.00000 0.9992
##
##              Class: 7 Class: 8 Class: 9 Class: 10
## Sensitivity          0.9900 0.9646 1.0000      NA
## Specificity          0.9930 1.0000 1.0000      1
## Pos Pred Value       0.9612 1.0000 1.0000      NA
## Neg Pred Value       0.9983 0.9929 1.0000      NA
## Prevalence           0.1481 0.1674 0.5259      0
## Detection Rate       0.1467 0.1615 0.5259      0
## Detection Prevalence 0.1526 0.1615 0.5259      0
## Balanced Accuracy     0.9915 0.9823 1.0000      NA

```

La tasa de precisión que alcanza este nuevo modelo basado en un árbol CART, establece lo adecuado que resulta para el problema que nos hemos planteado.

## 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 heteroscedasticidad, 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 para uno de los estratos de una de las variables, además de que para algunas categorías de la variable de respuesta se aprecia un desbalanceo en lo referente a la disponibilidad de casos etiquetados para algunas de las categorías:

Particularmente, el modelo para la calificación del nivel de la Contaminación del Aire que hemos obtenido, para la predicción del nivel de contaminación producido 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.