CO2_Emissions_Canada

Autor: Mikel Alvarez Rua y Mikel Tobar del Barrio

8 de Junio de 2021

1 Carga de librerías

En primer lugar, se obtienen las librerías necesarias para las operaciones a realizar.

```
#install.packages("ggplot2")
library("ggplot2")
#install.packages("gridExtra")
library("gridExtra")
#install.packages("fastDummies)
library("fastDummies")
#install.packages("kableExtra2")
library("kableExtra")
#install.packages("boruta")
library("Boruta")
#install.packages("ggcorrplot")
library("ggcorrplot")
#install.packages("pROC")
library("pROC")
#install.packages("tinytex")
library("tinytex")
#install.packages("webshot")
#webshot::install_phantomjs()
```

2 Descripción del dataset

Para la realización de esta práctica, se ha escogido el dataset `CO2 Emissions_Canada.csv` . El conjunto de datos recoge observaciones de un total de 7 años de una serie de modelos de coches diferentes, y da información acerca de las características técnicas de dichos coches y de su consumo en diversas situaciones.

El dataset procede del portal de datos abiertos del gobierno federal de Canadá, aunque se ha empleado una versión compilada de Kaggle. Se dan a continuación los enlaces de interés para el acceso a los datos:

- Dataset en el Open Data Canada: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64 (https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64)
- Enlace del dataset en Kaggle: https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles?select=Data+Description.csv (https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles?select=Data+Description.csv)

La licencia del dataset publicado en Kaggle es “Open Database License”, por lo que se puede compartir, modificar y usar libremente.

El dataset `CO2 Emissions_Canada.csv` está compuesto por 7385 observaciones de 12 variables. Las descripciones de algunas de las variables se encuentran en el archivo `Data Description.csv` :

```
desc<-read.csv("Data Description.csv")
names(desc)<-c("Variable", "Descripción del valor")
kable_styling(kable(desc, format='html', caption = "Descripcion valores de CO2 Emissio
ns_Canada"))
```

Descripcion valores de CO2 Emissions_Canada

Variable	**Descripción del valor**
Model	4WD/4X4 = Four-wheel drive
	AWD = All-wheel drive
	FFV = Flexible-fuel vehicle
	SWB = Short wheelbase
	LWB = Long wheelbase
	EWB = Extended wheelbase
Transmission	A = automatic
	AM = automated manual
	AS = automatic with select shift
e	AV = continuously variable
	M = manual
	3 - 10 = Number of gears
Fuel type	X = regular gasoline

Variable	Descripción del valor
	Z = premium gasoline
	D = diesel
	E = ethanol (E85)
	N = natural gas
Fuel consumption	City and highway fuel consumption ratings are shown in litres per 100 kilometres (L/100 km) - the combined rating (55% city, 45% hwy) is shown in L/100 km and in miles per imperial gallon (mpg)
CO2 emissions	the tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving

El resto de las variables son:

- `Make` : Marca del vehículo
- `Vehicle.Class` : Modelo del vehículo
- `Engine.Size.L.` : Cilindrada del motor en Litros
- `Cylinders` : El número de cilindros del motor

A partir de las variables anteriormente descritas, se pueden plantear las siguientes preguntas de investigación:

- ¿Cuáles son las correlaciones entre las diferentes variables que componen nuestro dataset?¿Qué variables están más estrechamente relacionadas entre sí?
- ¿Se puede afirmar que los coches manuales contaminan menos que el resto?
- ¿Cuáles son las variables que hacen que un coche sea menos contaminante?

3 Integración y selección

En primer lugar, se cargarán las librerías necesarias para la ejecución del código.

Leemos el archivo y observamos los tipos inferidos por R.

```
# Cargamos el juego de datos
co2<-read.csv("CO2 Emissions_Canada.csv")
str(co2)
```

```
## 'data.frame':    7385 obs. of  12 variables:
##  $ Make                           : chr  "ACURA" "ACURA" "ACURA" "ACURA" ...
##  $ Model                          : chr  "ILX" "ILX" "ILX HYBRID" "MDX 4WD" ...
##  $ Vehicle.Class                  : chr  "COMPACT" "COMPACT" "COMPACT" "SUV - SMAL
L" ...
##  $ Engine.Size.L.                 : num  2 2.4 1.5 3.5 3.5 3.5 3.5 3.7 3.7 2.4 ...
##  $ Cylinders                      : int  4 4 4 6 6 6 6 6 6 4 ...
##  $ Transmission                   : chr  "AS5" "M6" "AV7" "AS6" ...
##  $ Fuel.Type                      : chr  "Z" "Z" "Z" "Z" ...
##  $ Fuel.Consumption.City..L.100.km.: num  9.9 11.2 6 12.7 12.1 11.9 11.8 12.8 13.4
10.6 ...
##  $ Fuel.Consumption.Hwy..L.100.km. : num  6.7 7.7 5.8 9.1 8.7 7.7 8.1 9 9.5 7.5 ...
##  $ Fuel.Consumption.Comb..L.100.km.: num  8.5 9.6 5.9 11.1 10.6 10 10.1 11.1 11.6
9.2 ...
##  $ Fuel.Consumption.Comb..mpg.     : int  33 29 48 25 27 28 28 25 24 31 ...
##  $ CO2.Emissions.g.km.             : int  196 221 136 255 244 230 232 255 267 212
...
```

Observamos que los valores numéricos han sido reconocidos como tal mientras que el resto podemos transformarlos en factores de acuerdo con la descripción que hemos obtenido de ellos:

```
chars<-colnames(co2[sapply(co2,is.character)])
co2[chars]<-lapply(co2[chars],factor)
summary(co2)
```

```
##             Make                Model             Vehicle.Class  Engine.Size.L.
##  FORD         : 628   F-150 FFV    :  32   SUV - SMALL   :1217   Min.   :0.90
##  CHEVROLET    : 588   F-150 FFV 4X4:  32   MID-SIZE      :1133   1st Qu.:2.00
##  BMW          : 527   MUSTANG      :  27   COMPACT       :1022   Median :3.00
##  MERCEDES-BENZ: 419   FOCUS FFV    :  24   SUV - STANDARD: 735   Mean   :3.16
##  PORSCHE      : 376   F-150        :  20   FULL-SIZE     : 639   3rd Qu.:3.70
##  TOYOTA       : 330   F-150 4X4    :  20   SUBCOMPACT    : 606   Max.   :8.40
##  (Other)      :4517   (Other)      :7230   (Other)       :2033
##    Cylinders       Transmission  Fuel.Type Fuel.Consumption.City..L.100.km.
##  Min.   : 3.000   AS6    :1324   D: 175    Min.   : 4.20
##  1st Qu.: 4.000   AS8    :1211   E: 370    1st Qu.:10.10
##  Median : 6.000   M6     : 901   N:   1    Median :12.10
##  Mean   : 5.615   A6     : 789   X:3637    Mean   :12.56
##  3rd Qu.: 6.000   A8     : 490   Z:3202    3rd Qu.:14.60
##  Max.   :16.000   AM7    : 445             Max.   :30.60
##                   (Other):2225
##  Fuel.Consumption.Hwy..L.100.km. Fuel.Consumption.Comb..L.100.km.
##  Min.   : 4.000                  Min.   : 4.10
##  1st Qu.: 7.500                  1st Qu.: 8.90
##  Median : 8.700                  Median :10.60
##  Mean   : 9.042                  Mean   :10.98
##  3rd Qu.:10.200                  3rd Qu.:12.60
##  Max.   :20.600                  Max.   :26.10
##
##  Fuel.Consumption.Comb..mpg. CO2.Emissions.g.km.
##  Min.   :11.00               Min.   : 96.0
##  1st Qu.:22.00               1st Qu.:208.0
##  Median :27.00               Median :246.0
##  Mean   :27.48               Mean   :250.6
##  3rd Qu.:32.00               3rd Qu.:288.0
##  Max.   :69.00               Max.   :522.0
##
```

Comparando la descripción de la variable `Transmission` con sus valores, observamos que el tipo de transmisión y el número de marchas están incluidas en la misma variable. Procedemos a separarlas para poder analizarlas más detalladamente más adelante.

```
co2$Gears<-as.numeric(gsub("[a-zA-Z]*","",co2$Transmission))
co2$Transmission<-gsub("[0-9]","",co2$Transmission)
```

Antes de adentrarnos en la preparación y análisis de los datos, acortamos los nombres de algunas variables:

```
names(co2)[8:12]<-c("City.L.100km", "Hwy.L.100km", "Comb.L.100km", "Comb.mpg", "CO2.g.
km")
```

Mostramos un extracto de los datos para comprobar que los cambios se han efectuado correctamente:

```
head(co2,5)
```

```
##    Make      Model Vehicle.Class Engine.Size.L. Cylinders Transmission
## 1 ACURA        ILX       COMPACT            2.0         4           AS
## 2 ACURA        ILX       COMPACT            2.4         4            M
## 3 ACURA ILX HYBRID       COMPACT            1.5         4           AV
## 4 ACURA    MDX 4WD   SUV - SMALL            3.5         6           AS
## 5 ACURA    RDX AWD   SUV - SMALL            3.5         6           AS
##   Fuel.Type City.L.100km Hwy.L.100km Comb.L.100km Comb.mpg CO2.g.km Gears
## 1         Z          9.9         6.7          8.5       33      196     5
## 2         Z         11.2         7.7          9.6       29      221     6
## 3         Z          6.0         5.8          5.9       48      136     7
## 4         Z         12.7         9.1         11.1       25      255     6
## 5         Z         12.1         8.7         10.6       27      244     6
```

4 Limpieza de los datos

4.1 Valores vacíos y duplicados

Empezamos por mirar si hay duplicados y/o valores vacíos en el juego de datos.

```
#Contamos duplicados
nrow(co2[duplicated(co2),])
```

```
## [1] 1103
```

```
#Comprobamos valores vacíos
colSums(is.na(co2))
```

```
##           Make          Model  Vehicle.Class Engine.Size.L.      Cylinders
##              0              0              0              0              0
##   Transmission      Fuel.Type   City.L.100km    Hwy.L.100km   Comb.L.100km
##              0              0              0              0              0
##       Comb.mpg       CO2.g.km          Gears
##              0              0            295
```

```
sum(colSums(co2==0))
```

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

```
sum(colSums(co2==""))
```

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

No parecen existir valores vacíos o perdidos más allá de no estar especificado el número de marchas en algunos modelos, pero sí encontramos duplicados. Un total de 1103 registros aparecen más de una vez. Procedemos a eliminarlos:

```
co2<-co2[!duplicated(co2),]
nrow(co2[duplicated(co2),])
```

```
## [1] 0
```

4.2 Valores extremos o *outliers*

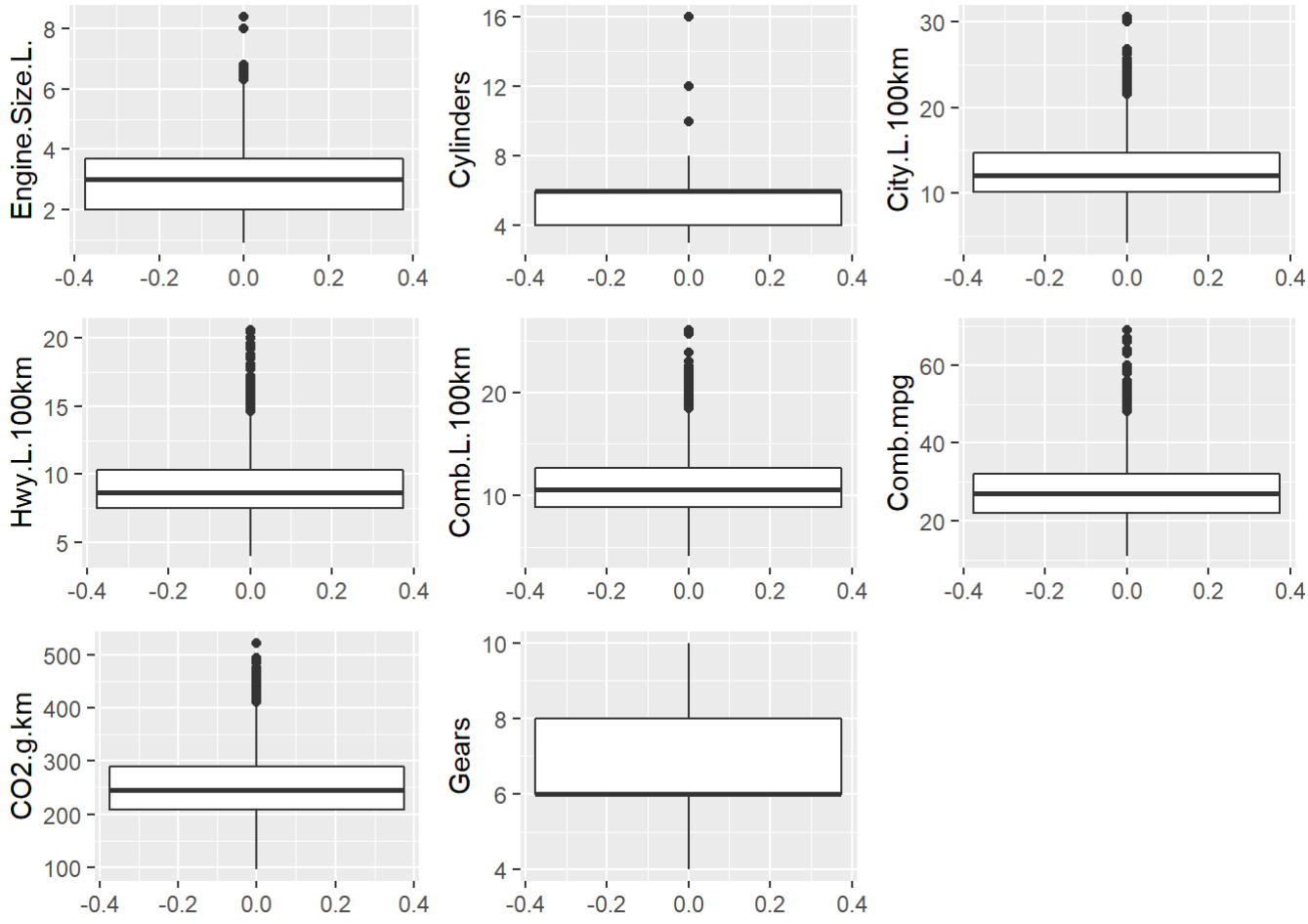
En el resumen, a simple vista, no se ha observado ningún valor máximo o mínimo muy alejado del resto de valores. Esto se aprecia mejor en gráficos de caja:

```
nums<-colnames(co2[sapply(co2,is.numeric)])

myplots <- vector('list', length(nums))

for (i in seq(1:length(nums))){
  message(i)
  myplots[[i]] <- local({
  i <- i
  h<-ggplot(co2, aes(y=.data[[nums[i]]])) +
  geom_boxplot()+scale_fill_brewer(palette="Dark2")+
  labs(y=nums[i])
})
}

do.call("grid.arrange", c(myplots, ncol=3))
```



Se efectúa a continuación el análisis de los diagramas de cajas y bigotes obtenidos.

- En el gráfico correspondiente al tamaño del motor, en litros (`Engine.Size.L`), se observa cómo existen algunos coches con un motor anormalmente grande. Sin embargo, se trata de motores que son utilizados por los modelos deportivos (Srt y Dodge Viper, Bugatti, etc.), por lo que serán valores que se incluirán en la muestra.
- En el gráfico correspondiente a los cilindros, se observa cómo algunos modelos disponen de una cantidad anormalmente elevada (16). Observando qué coches tienen este diseño, se encuentra lo mismo que en el apartado anterior: modelos deportivos con un diseño especial para alcanzar grandes velocidades. Por tanto, se aceptan los valores extremos.
- En el gráfico correspondiente al consumo en ciudad, se observa cómo algunos modelos sobresalen por la parte alta del gráfico. Se trata de furgonetas de consumo elevado, que se explica por su diseño y por su capacidad de carga. Se aceptan los valores extremos. También se da esta situación, y se llega a la misma conclusión, en los gráficos correspondientes al consumo en autopista y al consumo general (litros).
- En el boxplot correspondiente al consumo en mpg, un mpg alto significa un consumo bajo. Por tanto, los valores extremos observados corresponden a vehículos híbridos o de dimensiones pequeñas que no consumen mucha gasolina. Se validan los valores.
- Finalmente, en el CO^2 emitido, se observan valores extremos por arriba, explicados por las elevadas emisiones de coches deportivos, furgonetas de carga, y SUVs. Se da por bueno el valor extremo.
- Todos los valores extremos se dan por la parte alta de los diagramas.

5 Análisis de los datos

A continuación, se va a realizar un análisis de los datos que se pretenden estudiar.

5.1 Selección de los grupos de datos a analizar/comparar

Volvamos a estudiar las preguntas que se han planteado al principio. Las pruebas que se pretenden realizar son:

- Contraste de hipótesis para el estudio de la contaminación de coches manuales vs la del resto de vehículos.
- Regresión logística para la relevancia de las variables en cuanto al condicionamiento de la contaminación de un vehículo.
- Análisis de correlaciones para estudiar las interferencias entre las diferentes variables.

Para el contraste de hipótesis, será necesario dividir la muestra en dos grupos, el de los coches manuales y el del resto de vehículos.

Para la regresión logística y el análisis de correlaciones, se puede estudiar la muestra en su totalidad.

Por tanto, se contará con tres grupos de análisis.

- Muestra completa.
- Muestra de los coches manuales.
- Muestra de los coches no manuales.

Se divide a continuación la muestra en dos grupos para el contraste de hipótesis. Interesará únicamente el atributo relativo a la contaminación, junto al que se mostr

```
manual_automatico <- co2[,c("Transmission", "CO2.g.km")]
comparacion_manual <- subset(co2 , Transmission == "M")
comparacion_resto <- subset(co2 , Transmission != "M")
comparacion_manual$manual<-"Manual"
comparacion_resto$manual<-"Resto"
```

5.2 Normalidad y homogeneidad de la varianza

Para los análisis estadísticos como el contraste de hipótesis, es importante realizar una comprobación de la normalidad y la homogeneidad de la varianza en la variable de trabajo.

En primer lugar, para la comprobación de la normalidad en este tipo de pruebas estadísticas, es frecuente asumirla según el principio del Teorema del Límite Central. Dicho teorema que afirma que en las muestras de más de 30 elementos van a tender siempre a la normalidad. Se comprueba que nuestros 2 grupos de comparación en el contraste de hipótesis van a tener una distribución normal y se comprueba el número de elementos de cada uno de ellos.

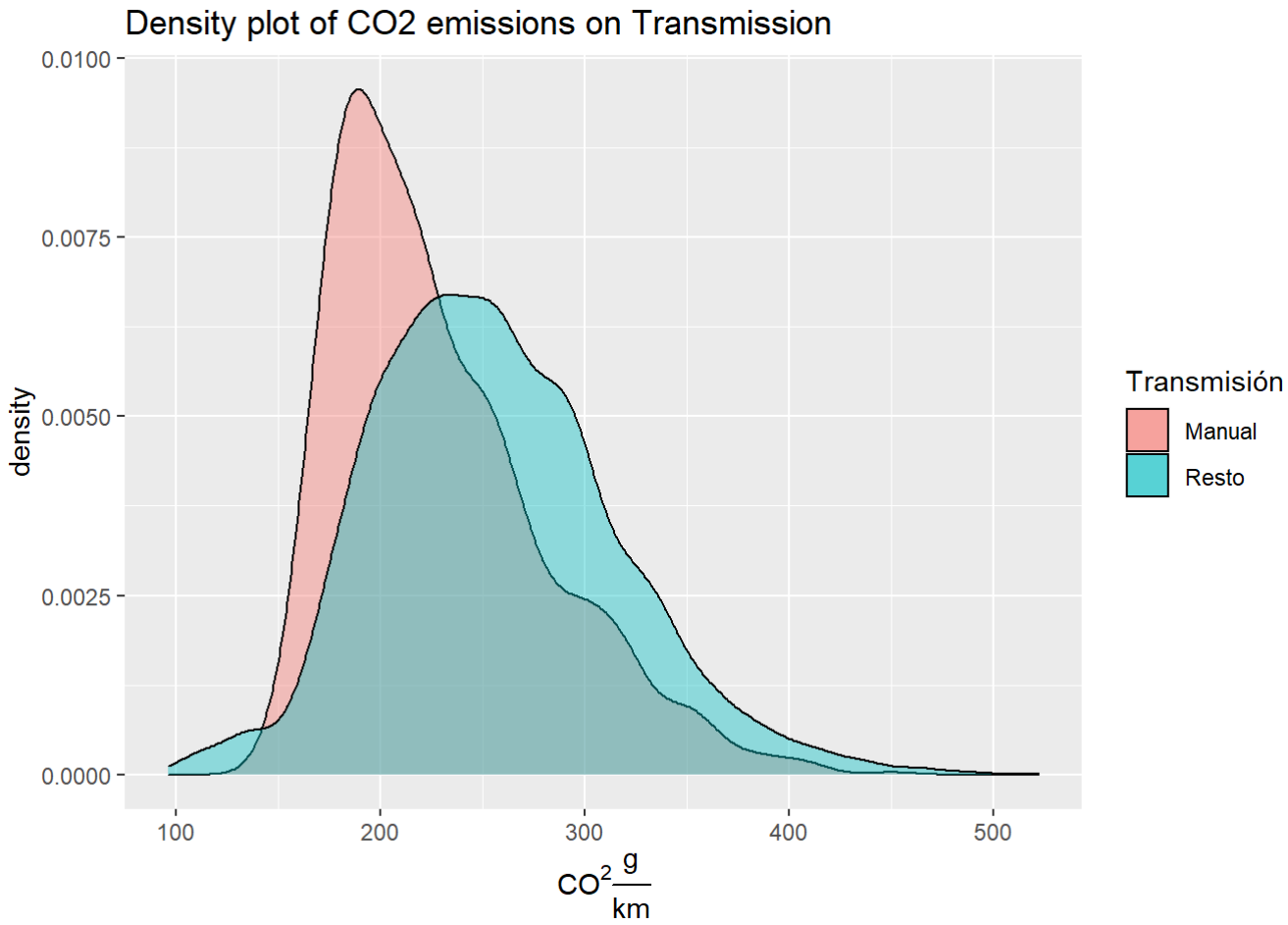
```
nrow(comparacion_manual)
```

```
## [1] 1019
```

```
nrow(comparacion_resto)
```

```
## [1] 5263
```

```
ggplot(data=comparacion_manual, aes(x=CO2.g.km, fill=manual))+ geom_density(alpha=0.4)
+
  ggtitle("Density plot of CO2 emissions on Transmission")+geom_density(data = compara
cion_resto, aes(x=CO2.g.km), alpha=0.4)+
  xlab(expression(paste(CO^{2},frac(g,km))))+labs(fill="Transmisión")
```



En segundo lugar, será importante hacer un test de varianzas de las dos muestras. Se trata de un test bilateral, con un intervalo de confianza del 95%, en el que la hipótesis nula es que las varianzas son iguales y la hipótesis alternativa es que son diferentes.

$$H_0 : \sigma_{manuales} = \sigma_{resto}$$

$$H_1 : \sigma_{manuales} \neq \sigma_{resto}$$

```
testVarianzas <- function(x,y){
  var.test(x,y)
}

testVarianzas(comparacion_manual$CO2.g.km, comparacion_resto$CO2.g.km)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.74222, num df = 1018, denom df = 5262, p-value = 2.861e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6760357 0.8174112
## sample estimates:
## ratio of variances
##          0.7422236
```

Se puede, por tanto, asumir heterocedasticidad (varianzas diferentes) por la vía del valor p, que es de prácticamente 0, por debajo de la significancia de $\alpha = 0.05$.

5.3 Pruebas estadísticas

A continuación, se llevan a cabo los métodos estadísticos que permiten realizar los cálculos que responderán a las preguntas que nos hemos realizado.

5.3.1 Contraste de hipótesis

Para el contraste de hipótesis, se ha de contar con una serie de elementos que definen este contraste.

1. Se trata de un contraste de dos muestras.
2. Se asume normalidad y heterocedasticidad.
3. Se emplearán tests paramétricos, asumiendo que las muestras siguen una normalidad y que los datos son suficientes.
4. Se realizarán tests unilaterales, ya que se quiere comparar si la contaminación de los coches manuales es significativamente menor que la del resto de coches, por tanto, analizando únicamente el segmento superior de la distribución.
5. Se asumirá que la varianza es desconocida, ya que, en el contexto de esta práctica, se trabaja con una muestra.
6. La hipótesis nula será que los coches manuales no contaminan menos que el resto, y la alternativa que sí lo hacen.

La fórmula del test estadístico elegido es la siguiente:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\upsilon} \qquad (14)$$

El test a emplear es el siguiente:

$$H_0 : \mu_{manuales} \geq \mu_{resto}$$

$$H_1 : \mu_{manuales} < \mu_{resto}$$

En primer lugar, se calculará el valor crítico asociado al intervalo de confianza del 95% con el que se trabajará.

```
qnorm(0.05)
```

```
## [1] -1.644854
```

Por tanto, este valor marcará la zona de aceptación de la hipótesis alternativa, que será (-∞, -1.64] y la zona de rechazo de la hipótesis alternativa (-1.64, ∞).

Se realiza a continuación el test estadístico correspondiente.

```
t.test(comparacion_manual$CO2.g.km, comparacion_resto$CO2.g.km,alternative=c("less"))
```

```
##
##  Welch Two Sample t-test
##
## data:  comparacion_manual$CO2.g.km and comparacion_resto$CO2.g.km
## t = -15.049, df = 1597.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -24.2492
## sample estimates:
## mean of x mean of y
##  228.3474  255.5742
```

En este caso, se puede afirmar que la hipótesis alternativa se cumple, al ser el valor p de prácticamente 0. El valor crítico da -15.05, muy por debajo del -1.64 establecido, y el valor p es menor a la significancia de 0.05. Se puede por tanto afirmar que **los coches manuales contaminan menos que los coches automáticos**.

Pero, podemos analizar qué variables más allá de la transmisión influyen en qué otras variables y en especial en la emisión de CO^2. Para ello estudiaremos las correlaciones entre variables:

5.3.2 Correlaciones

Analizaremos de entre todas las variables. Para ello es preciso transformar cada variable categórica en un conjunto de variables dicotómicas que adopten el valor 1 o TRUE cuando corresponda.

```
results <- dummy_cols(co2, select_columns = c("Transmission", "Fuel.Type"))
results$Fuel.Type<-NULL
```

```
nums2<-results[,colnames(results[sapply(results,is.numeric)])]
nums2$Gears<-NULL
source("http://www.sthda.com/upload/rquery_cormat.r")
cormat<-rquery.cormat(nums2, type="flatten", graph = FALSE)
```

```
## Warning: package 'corrplot' was built under R version 4.0.5
```

```
## corrplot 0.88 loaded
```

```
cormat.ordered<-head(cormat$r[order(abs(cormat$r$cor), decreasing = TRUE),],20)
kable_styling(kable(cormat.ordered, format='html', caption = "Correlaciones entre algu
nas variables"))
```

Correlaciones entre algunas variables

	row	column	cor	p
28	City.L.100km	Comb.L.100km	0.99	0
27	Hwy.L.100km	Comb.L.100km	0.98	0

	row	column	cor	p
21	Hwy.L.100km	City.L.100km	0.95	0
6	Engine.Size.L.	Cylinders	0.93	0
112	City.L.100km	Comb.mpg	-0.93	0
113	Comb.L.100km	Comb.mpg	-0.93	0
20	CO2.g.km	City.L.100km	0.92	0
26	CO2.g.km	Comb.L.100km	0.92	0
110	CO2.g.km	Comb.mpg	-0.91	0
111	Hwy.L.100km	Comb.mpg	-0.89	0
15	CO2.g.km	Hwy.L.100km	0.88	0
93	Fuel.Type_Z	Fuel.Type_X	-0.86	0
9	Engine.Size.L.	CO2.g.km	0.85	0
10	Cylinders	CO2.g.km	0.83	0
18	Engine.Size.L.	City.L.100km	0.83	0
24	Engine.Size.L.	Comb.L.100km	0.82	0
19	Cylinders	City.L.100km	0.80	0
25	Cylinders	Comb.L.100km	0.78	0
13	Engine.Size.L.	Hwy.L.100km	0.77	0
108	Engine.Size.L.	Comb.mpg	-0.76	0

Observamos en la tabla un hecho intuitivo: el consumo de combustible está muy estrechamente relacionado con la cantidad de CO^2 emitido.

Para tener una idea más detallada de los factores que influyen en las emisiones del vehículos creamos un modelo de regresión logística.

5.3.3 Regresión logística

Se puede realizar una regresión logística para estudiar la contaminación producida por los coches.

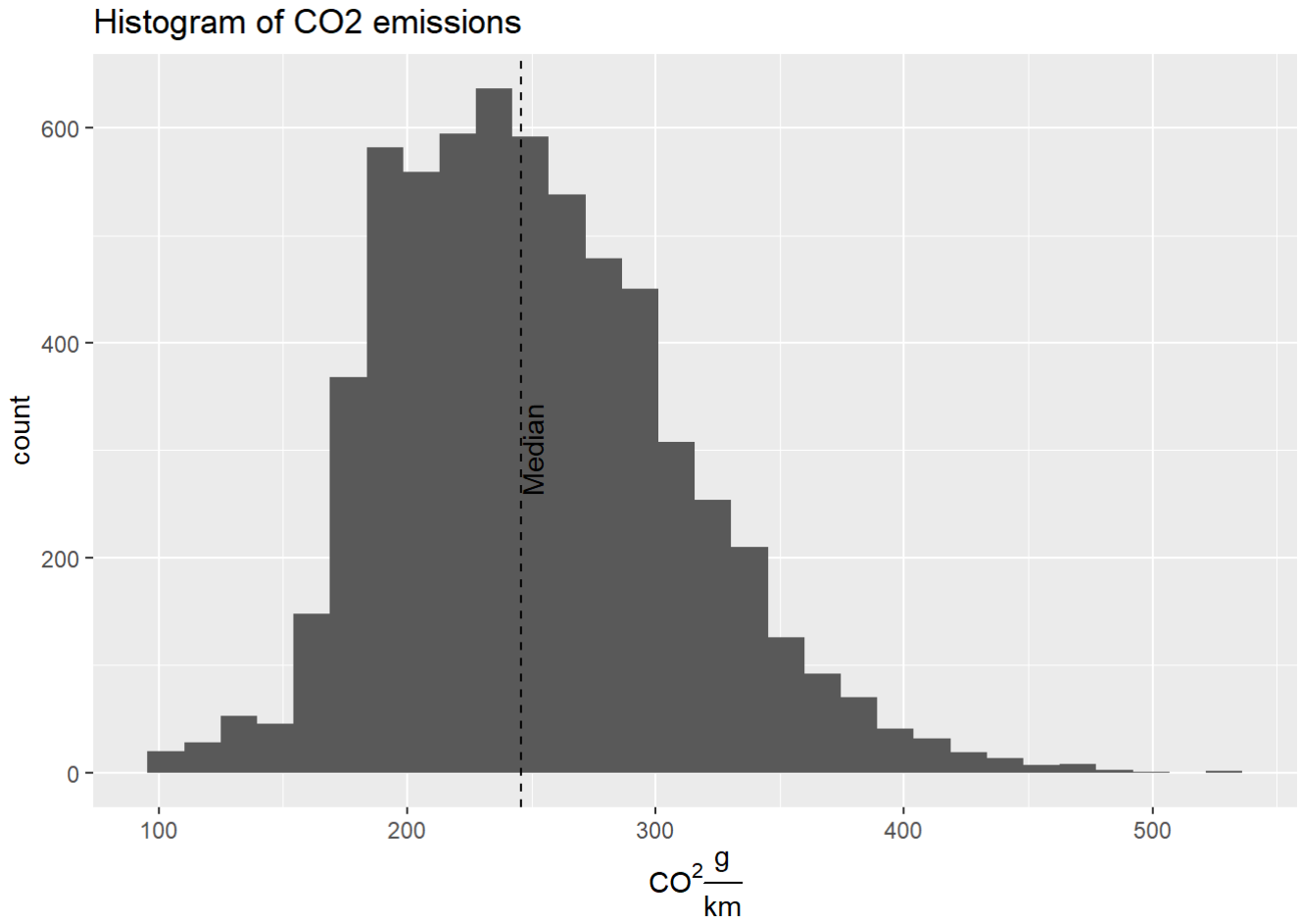
Para este apartado, se considera importante disponer de una variable descriptora. Se pretende enfocar el dataset hacia el estudio de cuánto CO^2 genera cada uno de los modelos por kilómetro, para observar su impacto medioambiental.`

Por tanto, se considera interesante observar cómo se distribuye la contaminación en la muestra que tenemos, mediante un análisis de sus medidas estadística, para poder trazar la frontera entre dos grupos en función de su contaminación, y poder binarizar la variable y obtener una nueva, que será aplicable a algunos de los modelos que se realizarán.

```
ggplot(data = co2, aes(x=CO2.g.km)) + geom_histogram() +
  annotate(geom = "vline",
             x = median(co2$CO2.g.km),
             xintercept = median(co2$CO2.g.km),
             linetype = "dashed")+
  annotate(geom = "text",
             label = "Median",
             x = median(co2$CO2.g.km),
             y = 300,
             angle = 90,
             vjust = 1)+
  ggtitle("Histogram of CO2 emissions")+
  xlab(expression(paste(CO^{2},frac(g,km))))
```

```
## Warning: Ignoring unknown aesthetics: x
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summary(co2$CO2.g.km)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    96.0   208.0   246.0   251.2   289.0   522.0
```

Se considera que se puede establecer la mediana, 246 g de CO^2 por km como valor fronterizo, al ser la medida central.

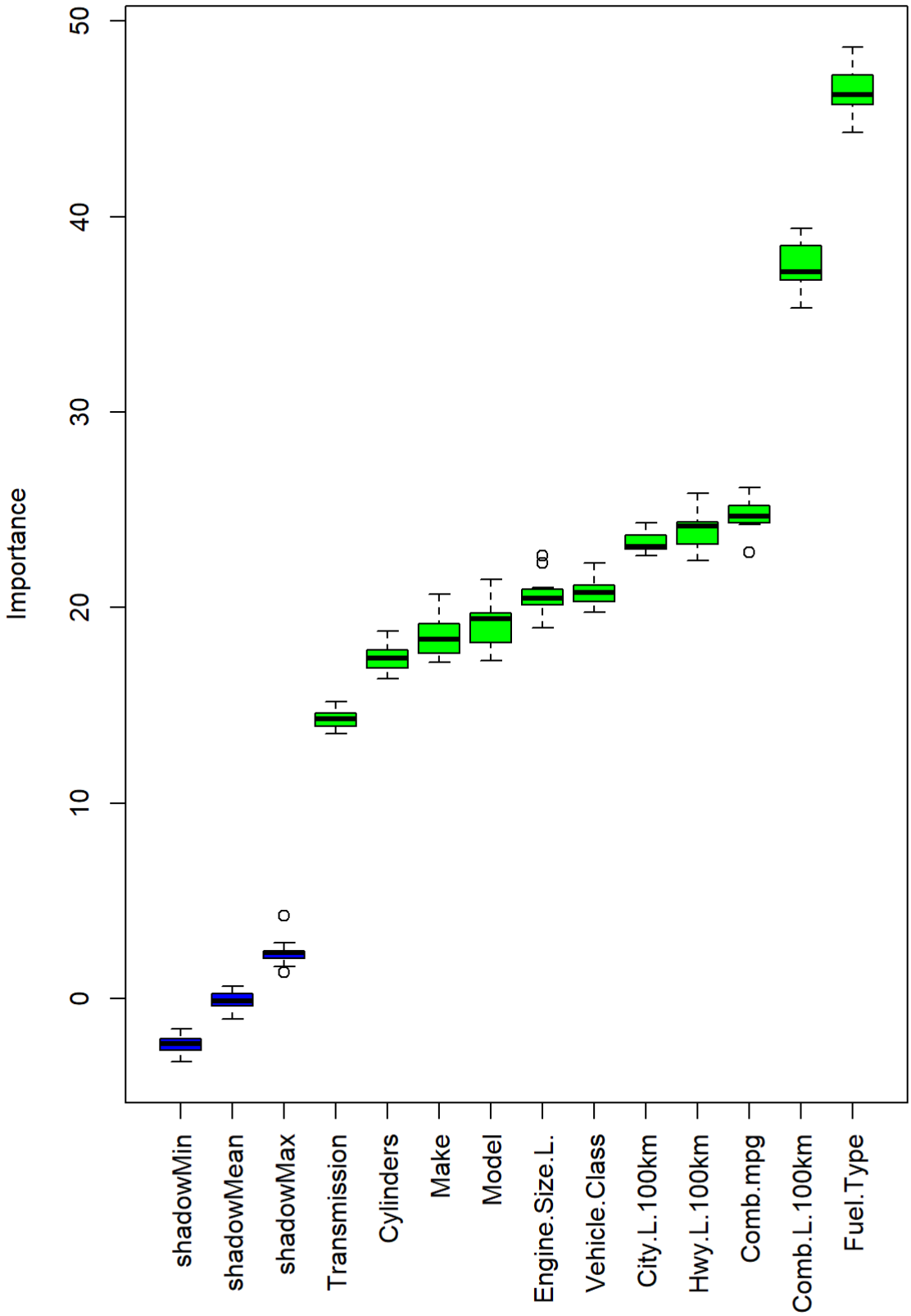
```
co2["co2.g.km.binary"] <- cut(co2$CO2.g.km, breaks = c(0,246,10000), labels = c("0",
"1"))
```

```
co2.boruta<-co2
co2.boruta$Gears<-NULL
co2.boruta$CO2.g.km<-NULL
boruta.co2 <- Boruta(co2.g.km.binary~., data = co2.boruta, doTrace = 2)
```

```
print(boruta.co2)
```

```
## Boruta performed 11 iterations in 12.56794 secs.
##  11 attributes confirmed important: City.L.100km, Comb.L.100km,
## Comb.mpg, Cylinders, Engine.Size.L. and 6 more;
##  No attributes deemed unimportant.
```

```
par(mar=c(10,5,5,5)+.1)
plot(boruta.co2, xlab= "", las=3)
```




```
#text(par("usr")[3] - 0.2,  srt = 45, pos = 1, xpd = TRUE)
```

A partir de aquí es posible plantear una regresión logística que modele la variable discretizada. Del modelo se excluye el modelo de vehículo para simplificarlo.

```
co2.boruta$Model<-NULL
glm.co2<- glm(co2.g.km.binary~., family=binomial, data=co2.boruta)
```

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

```
summary(glm.co2)
```

```
##
## Call:
## glm(formula = co2.g.km.binary ~ ., family = binomial, data = co2.boruta)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.724   0.000   0.000   0.000   2.382
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -4.515e+02  6.195e+01  -7.288 3.15e-13
## MakeALFA ROMEO                    6.080e+00  3.506e+04   0.000  0.99986
## MakeASTON MARTIN                  2.734e+00  4.193e+03   0.001  0.99948
## MakeAUDI                          1.271e+00  1.652e+00   0.769  0.44187
## MakeBENTLEY                      -3.475e+01  4.300e+03  -0.008  0.99355
## MakeBMW                           5.962e-01  1.580e+00   0.377  0.70598
## MakeBUGATTI                      -3.560e+02  3.408e+04  -0.010  0.99167
## MakeBUICK                         1.060e+00  1.789e+00   0.592  0.55356
## MakeCADILLAC                      8.415e-01  1.939e+00   0.434  0.66430
## MakeCHEVROLET                     5.183e-01  1.985e+00   0.261  0.79406
## MakeCHRYSLER                      1.037e-01  2.284e+00   0.045  0.96377
## MakeDODGE                        -5.459e-01  4.485e+00  -0.122  0.90312
## MakeFIAT                          9.330e+00  3.836e+03   0.002  0.99806
## MakeFORD                         -2.033e+00  2.046e+00  -0.994  0.32036
## MakeGENESIS                       1.488e+00  4.846e+01   0.031  0.97551
## MakeGMC                           2.245e+00  2.440e+00   0.920  0.35753
## MakeHONDA                        -6.197e-01  2.096e+00  -0.296  0.76751
## MakeHYUNDAI                      -1.124e+00  2.120e+00  -0.530  0.59591
## MakeINFINITI                      1.937e+00  2.148e+00   0.902  0.36713
## MakeJAGUAR                       -1.212e+00  2.590e+00  -0.468  0.63981
## MakeJEEP                          1.817e+00  2.167e+00   0.838  0.40178
## MakeKIA                          -1.081e+00  2.042e+00  -0.529  0.59650
## MakeLAMBORGHINI                  -9.085e+01  4.058e+03  -0.022  0.98214
## MakeLAND ROVER                    8.315e-01  1.980e+00   0.420  0.67456
## MakeLEXUS                        -2.596e+00  1.659e+00  -1.565  0.11767
## MakeLINCOLN                      -8.620e-01  1.056e+01  -0.082  0.93494
## MakeMASERATI                     -2.532e+01  4.300e+03  -0.006  0.99530
## MakeMAZDA                        -1.470e+01  2.187e+03  -0.007  0.99464
## MakeMERCEDES-BENZ                -7.094e-01  1.687e+00  -0.421  0.67403
## MakeMINI                          2.123e+00  1.857e+03   0.001  0.99909
## MakeMITSUBISHI                   -1.300e+00  1.856e+00  -0.700  0.48369
## MakeNISSAN                        2.395e+00  2.925e+00   0.819  0.41297
## MakePORSCHE                       2.168e+00  2.272e+00   0.954  0.33998
## MakeRAM                          -6.722e-01  8.207e+00  -0.082  0.93472
## MakeROLLS-ROYCE                  -1.333e+02  4.419e+03  -0.030  0.97594
## MakeSCION                        -2.193e+00  7.415e+03   0.000  0.99976
## MakeSMART                         6.479e+01  1.652e+04   0.004  0.99687
## MakeSRT                          -1.687e+02  3.408e+04  -0.005  0.99605
## MakeSUBARU                        5.466e+00  3.004e+00   1.819  0.06887
## MakeTOYOTA                       -2.969e+00  2.107e+00  -1.409  0.15871
## MakeVOLKSWAGEN                   -1.478e+00  1.630e+00  -0.907  0.36455
```

```
## MakeVOLVO                                2.854e+00  3.039e+00   0.939  0.34771
## Vehicle.ClassFULL-SIZE                  -1.764e+00  9.978e-01  -1.768  0.07702
## Vehicle.ClassMID-SIZE                   -7.105e-01  8.135e-01  -0.873  0.38246
## Vehicle.ClassMINICOMPACT                -4.230e+00  1.848e+00  -2.289  0.02210
## Vehicle.ClassMINIVAN                     1.549e+00  1.430e+00   1.083  0.27883
## Vehicle.ClassPICKUP TRUCK - SMALL        1.455e+00  1.487e+00   0.979  0.32774
## Vehicle.ClassPICKUP TRUCK - STANDARD     1.174e-01  1.730e+00   0.068  0.94590
## Vehicle.ClassSPECIAL PURPOSE VEHICLE     3.905e+00  1.506e+00   2.592  0.00953
## Vehicle.ClassSTATION WAGON - MID-SIZE    1.152e-02  2.785e+00   0.004  0.99670
## Vehicle.ClassSTATION WAGON - SMALL      -1.036e+00  3.096e+00  -0.335  0.73798
## Vehicle.ClassSUBCOMPACT                 -5.745e-01  7.373e-01  -0.779  0.43589
## Vehicle.ClassSUV - SMALL                -4.282e-01  9.830e-01  -0.436  0.66311
## Vehicle.ClassSUV - STANDARD              6.031e-01  1.160e+00   0.520  0.60311
## Vehicle.ClassTWO-SEATER                 -1.824e+00  1.690e+00  -1.079  0.28045
## Vehicle.ClassVAN - CARGO                -1.196e+02  6.709e+03  -0.018  0.98578
## Vehicle.ClassVAN - PASSENGER            -9.845e+01  2.870e+03  -0.034  0.97264
## Engine.Size.L.                          -1.137e+00  9.320e-01  -1.219  0.22267
## Cylinders                                1.055e+00  6.272e-01   1.683  0.09244
## TransmissionAM                           2.450e+00  1.096e+00   2.235  0.02540
## TransmissionAS                           1.796e+00  6.433e-01   2.791  0.00525
## TransmissionAV                          -2.429e+00  2.197e+00  -1.106  0.26875
## TransmissionM                            1.697e+00  1.004e+00   1.691  0.09082
## Fuel.TypeE                              -1.608e+02  1.778e+01  -9.047  < 2e-16
## Fuel.TypeN                              -1.148e+02  4.820e+04  -0.002  0.99810
## Fuel.TypeX                              -3.313e+01  3.835e+00  -8.639  < 2e-16
## Fuel.TypeZ                              -3.460e+01  3.992e+00  -8.669  < 2e-16
## City.L.100km                             9.584e+00  3.173e+00   3.020  0.00252
## Hwy.L.100km                              9.698e+00  2.803e+00   3.460  0.00054
## Comb.L.100km                             1.769e+01  6.133e+00   2.884  0.00393
## Comb.mpg                                 3.519e+00  7.544e-01   4.665 3.08e-06
##
## (Intercept)                             ***
## MakeALFA ROMEO
## MakeASTON MARTIN
## MakeAUDI
## MakeBENTLEY
## MakeBMW
## MakeBUGATTI
## MakeBUICK
## MakeCADILLAC
## MakeCHEVROLET
## MakeCHRYSLER
## MakeDODGE
## MakeFIAT
## MakeFORD
## MakeGENESIS
## MakeGMC
## MakeHONDA
## MakeHYUNDAI
## MakeINFINITI
## MakeJAGUAR
## MakeJEEP
```

```
## MakeKIA
## MakeLAMBORGHINI
## MakeLAND ROVER
## MakeLEXUS
## MakeLINCOLN
## MakeMASERATI
## MakeMAZDA
## MakeMERCEDES-BENZ
## MakeMINI
## MakeMITSUBISHI
## MakeNISSAN
## MakePORSCHE
## MakeRAM
## MakeROLLS-ROYCE
## MakeSCION
## MakeSMART
## MakeSRT
## MakeSUBARU                              .
## MakeTOYOTA
## MakeVOLKSWAGEN
## MakeVOLVO
## Vehicle.ClassFULL-SIZE                  .
## Vehicle.ClassMID-SIZE
## Vehicle.ClassMINICOMPACT                *
## Vehicle.ClassMINIVAN
## Vehicle.ClassPICKUP TRUCK - SMALL
## Vehicle.ClassPICKUP TRUCK - STANDARD
## Vehicle.ClassSPECIAL PURPOSE VEHICLE  **
## Vehicle.ClassSTATION WAGON - MID-SIZE
## Vehicle.ClassSTATION WAGON - SMALL
## Vehicle.ClassSUBCOMPACT
## Vehicle.ClassSUV - SMALL
## Vehicle.ClassSUV - STANDARD
## Vehicle.ClassTWO-SEATER
## Vehicle.ClassVAN - CARGO
## Vehicle.ClassVAN - PASSENGER
## Engine.Size.L.
## Cylinders                               .
## TransmissionAM                          *
## TransmissionAS                          **
## TransmissionAV
## TransmissionM                           .
## Fuel.TypeE                              ***
## Fuel.TypeN
## Fuel.TypeX                              ***
## Fuel.TypeZ                              ***
## City.L.100km                            **
## Hwy.L.100km                             ***
## Comb.L.100km                            **
## Comb.mpg                                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8706.48  on 6281  degrees of freedom
## Residual deviance:  271.51  on 6211  degrees of freedom
## AIC: 413.51
##
## Number of Fisher Scoring iterations: 21
```

Se aprecia que tanto el análisis llevado a cabo por Boruta como el modelo de regresión logística generado coinciden en marcar algunas variables como de mayor significancia. Estas son: el tipo de combustible y el consumo.

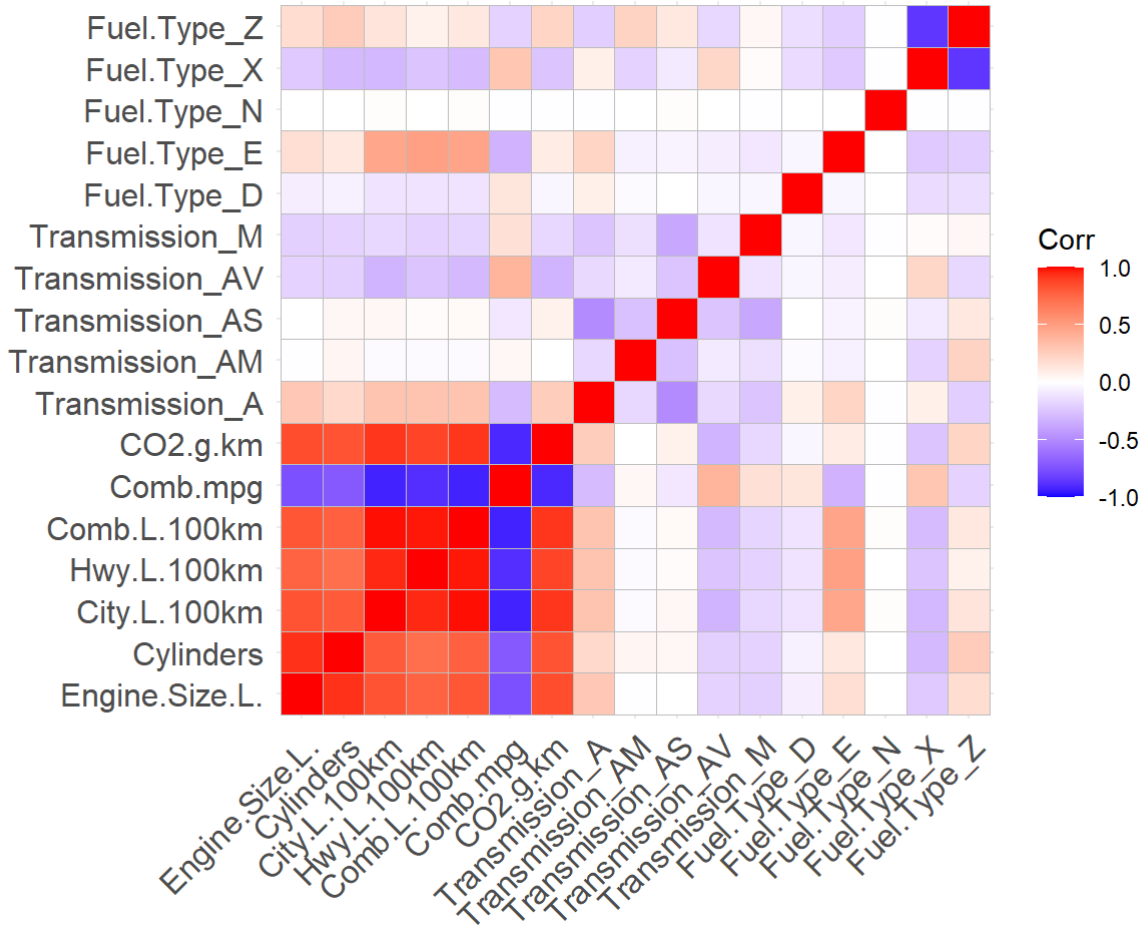
Se observan discrepancias entre ambos análisis en cuanto a las variables de tipo de transmisión y cilindrada. Mientras que hay variables que no influyen tanto, como la marca y el modelo de coche.

6 Representación gráfica

Una manera más intuitiva de observar los resultados de los análisis anteriores es por medio de gráficos. A continuación, se muestran de forma gráfica los resultados obtenidos anteriormente:

6.1 Matriz de correlaciones

```
corrs <- round(cor(nums2), 2)
ggcorrplot(corrs)
```



Además de las obersvaciones anteriores, podemo destacar viendo la matriz de correlaciones que los tipos de transmisión tienen correlaciones negativas entre ellos como es de esperar debido a que hay coches que solo están disponibles con un tipo de transmisión y por lo tanto el que haya un tipo excluye al resto.

También observamos, que hay algunos tipos de combustible como son el Diesel `D` y la gasolina común `X` que tienen una correlación positiva con el consumo y por tanto con la emisión de CO^2; mientras que el Ethanol (E85) `E` y la gasolina premium `Z` tienen una correlación negativa.

6.2 Análisis de la bondad de ajuste de la regresión

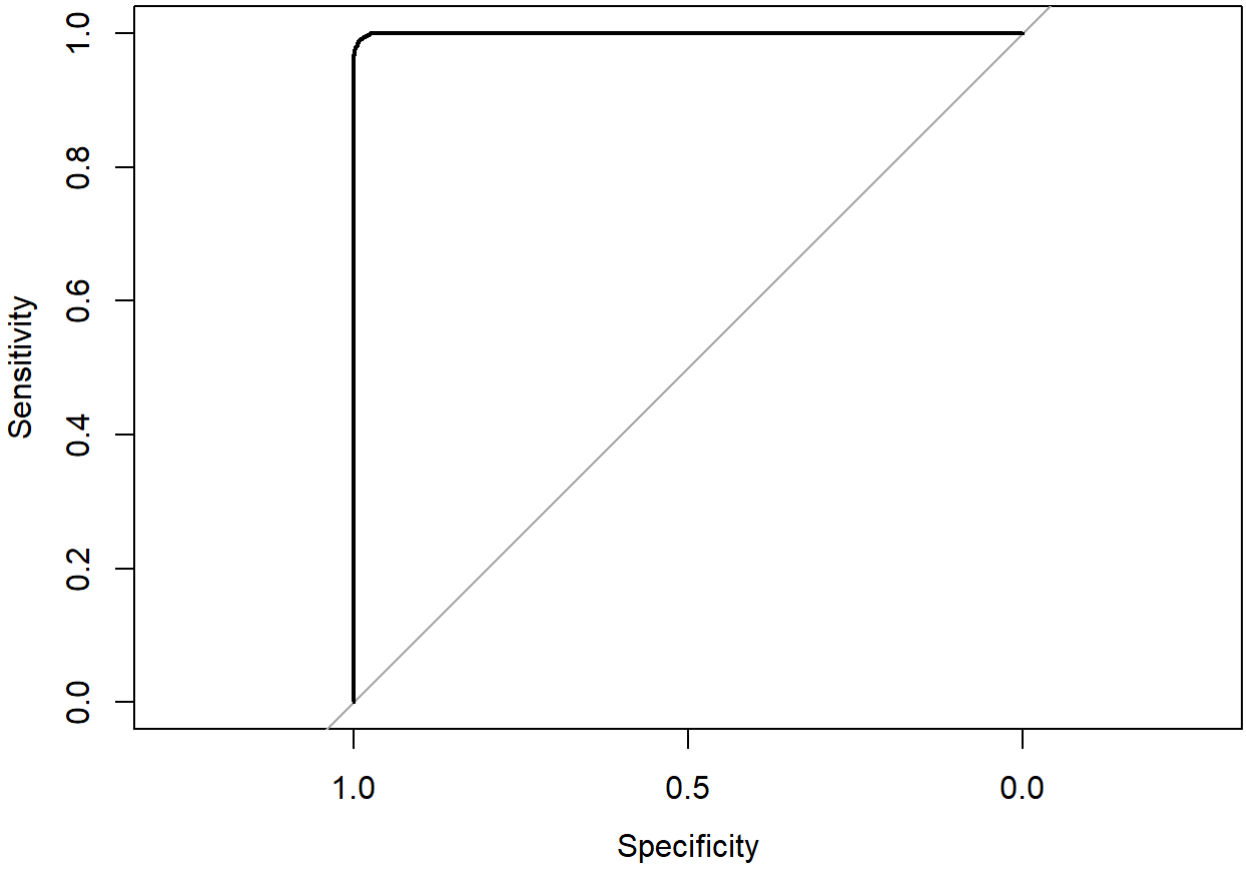
Para estimar la bondad de ajuste del modelo de regresión creado podemos observar la curva de ROC.

```
p1=predict(glm.co2, co2.boruta, type="response")
r1=roc(co2.boruta$co2.g.km.binary,p1, data=co2.boruta)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(r1)
```




```
auc(r1)
```

```
## Area under the curve: 0.9997
```

Con un área bajo la curva de 0.9997, el modelo de regresión logística se puede considerar con muy buen ajuste. Esto se debe a que la emisión de CO^2 y el consumo de combustible están muy intimamente relacionados como hemos visto también en el análisis de correlaciones.

7 Conclusiones

A continuación, se indican las conclusiones principales que ha arrojado el análisis del dataset:

- Si bien el contraste de hipótesis arroja que los coches no manuales contaminan más en general, el análisis vía Boruta no la sitúa entre las variables más importantes.
- El estudio de correlaciones, el análisis Boruta y la regresión logística dan gran importancia a las variables "Fuel.Type" y en general todas las relacionadas con el consumo. Son dos variables que a priori, se hubieran indicado como importantes para el análisis.
- Sin embargo, los análisis realizados no acuerdan tanta importancia a variables como la marca y el modelo no influyen.

8 Tabla de contribuciones

Contribuciones	*Firma*
Investigación previa	
Redacción de las respuestas	
Desarrollo código