Seleccion de un conjunto de datos adicional

De Fino - Solari Barrios - Wurzel

2023-08-18

1.

Los datos se obtuvieron del sitio Kaggle: https://www.kaggle.com/datasets/bhuviranga/co2-emissions

```
df = read.csv('./data/CO2_Emissions.csv')
head(df,5)
```

```
##    Make      Model Vehicle.Class Engine.Size.L. Cylinders Transmission
## 1 ACURA        ILX       COMPACT            2.0         4          AS5
## 2 ACURA        ILX       COMPACT            2.4         4           M6
## 3 ACURA ILX HYBRID       COMPACT            1.5         4          AV7
## 4 ACURA    MDX 4WD   SUV - SMALL            3.5         6          AS6
## 5 ACURA    RDX AWD   SUV - SMALL            3.5         6          AS6
##   Fuel.Type Fuel.Consumption.City..L.100.km. Fuel.Consumption.Hwy..L.100.km.
## 1         Z                             9.9                            6.7
## 2         Z                            11.2                            7.7
## 3         Z                             6.0                            5.8
## 4         Z                            12.7                            9.1
## 5         Z                            12.1                            8.7
##   Fuel.Consumption.Comb..L.100.km. Fuel.Consumption.Comb..mpg.
## 1                              8.5                          33
## 2                              9.6                          29
## 3                              5.9                          48
## 4                             11.1                          25
## 5                             10.6                          27
##   CO2.Emissions.g.km
## 1                196
## 2                221
## 3                136
## 4                255
## 5                244
```

2.

```
#Summary de las cosas del dataset
str(df)
```

```
## '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 - SMALL" ...
##  $ 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 ...
```

La cantidad de observaciones de mi dataset es : 7385

La cantidad de variables es : 12

Pasamos a realizar una pequeña descripción de todas las variables:

- Make: fabricante del auto
- Model: modelo del auto
- Vehicle.Class: carrocería
- Engine.Size.L.: tamaño del motor en litros
- Cylinders: cantidad de cilindros
- Transmission: Transmision
- Fuel.Type: tipo de combustible
- Fuel.Consumption.City..L.100.km.: consumision de combustible en la ciudad medido en litros sobre 100km
- Fuel.Consumption.Hwy..L.100.km.: consumision de combustible en ruta medido en litros sobre 100km
- Fuel.Consumption.Comb..L.100.km.: consumision de combustible en ciudad y ruta combinados medido en litros por 100km
- Fuel.Consumption.Comb..mpg.: consumision de combustible en ciudad y ruta combinados medido en millas por galón
- CO2.Emissions.g.km: emisiones de carbono medido en gramos por kilometro

```
#Librería que vamos a utilizar para analizar mi dataset
suppressPackageStartupMessages(library("dplyr"))
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
#Tipo de variables:
length(select_if(df,is.numeric))
```

```
## [1] 7
```

```
length(select_if(df,is.logical))
```

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

```
length(select_if(df,is.character))
```

```
## [1] 5
```

```
#Cantidad de valores faltantes:
sum(sapply(df, function(x) sum(is.na(x))))
```

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

Mi dataset contiene:

- 7 variables numéricas
- 0 variables lógicas
- 5 variables categóricas

Por otro lado, mi dataset tiene una cantidad total de cero valores NAs

3.

Se intentará predecir si las emisiones de carbono del automóvil superarán 200 g/km, calificando para la mayor tasa de impuestos por emisiones (14,75%) según la normativa española IEDMT (Impuesto Esecial sobre Determinados Medios de Transporte, Artículo 65 de la Ley 38/1992). Elegimos establecer el umbral propuesto por España al ser el país cuya normativa fue más clara. Argentina por el momento no tiene una regulación vigente para vehículos particulares.

4.

Decidimos realizar una única transformación al conjunto de datos modificando la columna "CO2.Emissions.g.km" donde 1 indica que el vehiculo califica para el impuesto (CO2.Emissions.g.km > 200 g/km) y 0 si no.

```
mayores_a_200gkm <- df$CO2.Emissions.g.km > 200
mayores_a_200gkm <- as.integer(as.logical(mayores_a_200gkm))

df$CO2.Emissions.g.km <- mayores_a_200gkm
```

5.

Creamos el nuevo cvs con los datos transformados:

```
write.csv(df, "./data/CO2_Emissions_Transformado.csv", row.names=FALSE)
```