

Seleccion de un conjunto de datos adicional

De Fino - Solari Barrios - Wurzel

2023-08-22

Fuente de los datos

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

```
df = read.csv('./data/CO2_Emissions.csv')
#Desplegamos los primeros 5 datos de todas las variables del dataset
head(df[0:5],5)
```

	Make	Model	Vehicle.Class	Engine.Size.L.	Cylinders
	<chr>	<chr>	<chr>	<dbl>	<int>
1	ACURA	ILX	COMPACT	2.0	4
2	ACURA	ILX	COMPACT	2.4	4
3	ACURA	ILX HYBRID	COMPACT	1.5	4
4	ACURA	MDX 4WD	SUV - SMALL	3.5	6
5	ACURA	RDX AWD	SUV - SMALL	3.5	6
5 rows					

```
head(df[6:8],5)
```

	Transmission	Fuel.Type	Fuel.Consumption.City..L.100.km.
	<chr>	<chr>	<dbl>
1	AS5	Z	9.9
2	M6	Z	11.2
3	AV7	Z	6.0
4	AS6	Z	12.7
5	AS6	Z	12.1
5 rows			

```
head(df[9:10],5)
```

	Fuel.Consumption.Hwy..L.100.km.	Fuel.Consumption.Comb..L.100.km.
	<dbl>	<dbl>
1	6.7	8.5

	Fuel.Consumption.Hwy..L.100.km.	Fuel.Consumption.Comb..L.100.km.
2	7.7	9.8
3	5.8	5.9
4	9.1	11.1
5	8.7	10.6

5 rows

```
head(df[11:12],5)
```

	Fuel.Consumption.Comb..mpg.	CO2.Emissions.g.km
	<int>	<int>
1	33	196
2	29	221
3	48	136
4	25	255
5	27	244

5 rows

Características principales

```
#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

Nuestro dataset no contiene NAs

Lo que vamos a predecir

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 Especial 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.

Transformación a los datos

Decidimos realizar una única transformación al conjunto de datos modificando la columna "CO2.Emissions.g.km" para que 1 indique si el vehículo califica para el impuesto ($\text{CO2.Emissions.g.km} > 200$ g/km) y 0 si no. Además renombramos la columna a "CO2.Emissions.gt.200gkm" (CO2 Emissions greater than 200gkm).

```
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

df <- df %>%
  rename(CO2.Emissions.gt.200gkm = CO2.Emissions.g.km)
```

Guardamos el nuevo archivo

Creamos el nuevo cvs con los datos transformados:

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