

## Tipología y Ciclo de Vida de los Datos – Práctica 1

David Durán Olivar y Pablo Martínez Pavón

### Contexto

Actualmente, uno de los principales gastos a los que se ven abocados las personas es la compra de un coche. Así, muchas veces, bien por disponibilidad financiera, bien por querer un modelo antiguo que ya no se comercializa o simplemente por el hecho de despreocuparse de tener que cuidar un coche nuevo; todos los años muchas personas deciden comprar un coche usado.

Las estadísticas demuestran que en España las ventas de coches semiusados son superiores a los de coches nuevos. En 2019, según datos de GANVAM (Asociación Nacional de Vendedores de Vehículos a Motor, Reparación y Recambios) y ANFAC (Asociación Española de Fabricantes de Automóviles y Camiones) se comercializaron un total de 2.236.406 vehículos usados contra 1.501.224 de vehículos nuevos matriculados.

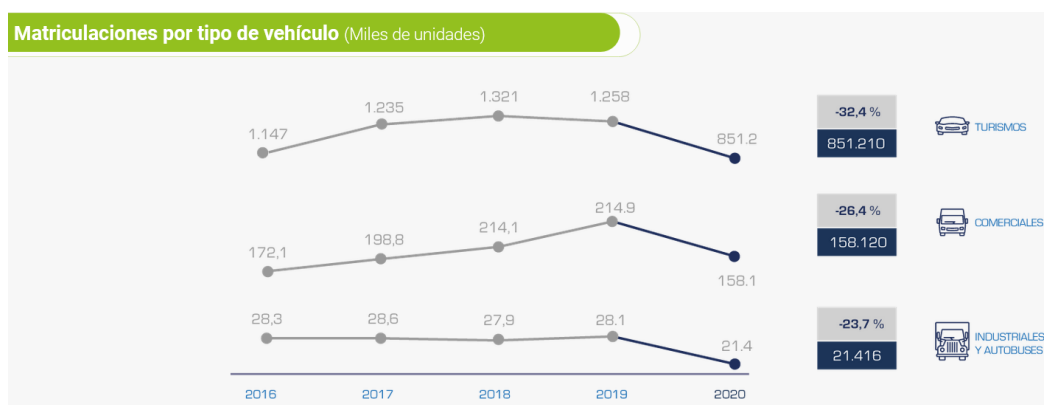


Figura 1. Evolución del número de matriculaciones de vehículo nuevo durante el último lustro.  
Fuente: ANFAC.

En el 2020, esta tendencia se dispara ya que con el COVID-19 a pesar de una disminución generalizada de las ventas tanto de coches usados, como de coche nuevos, la gente parece haber seguido optando de manera generalizada por los coches usados: 1.963.053 contra 1.030.746.

<https://www.ganvam.es/wp-content/uploads/2021/01/07-01-21-NP-VO-2020.pdf>

<https://anfacs.com/datos-clave-del-sector-automocion-2020/>

Así, una herramienta que analice los anuncios de vehículos de segunda mano puede resultar útil a muchas personas para elegir el anuncio correcto y reducir el gasto de su compra.

En los últimos años hemos asistido a un proceso de transformación social que busca fomentar un sistema y una sociedad más sostenible y respetuoso con el medio ambiente al apostar por soluciones que busquen la descarbonización y la economía circular. El sector de automoción, como respuesta a esta tendencia ha apostado por el desarrollo de vehículos híbridos o ecológicos que año tras año aumentan su porcentaje sobre el total de las ventas de vehículo nuevo. Según los datos de ANFAC entre 2020 y 2019 se ha producido un incremento del 26,2% de las matriculaciones de vehículos de movilidad alternativa en un contexto de pandemia en el que el conjunto de las ventas ha descendido.

Con nuestra herramienta también se podría comprobar si esta tendencia se ha trasladado a las ventas de vehículos usados.

## Título

CarAds

## Descripción del dataset

El dataset CarAds contiene una serie de registros de anuncios de vehículos de segunda mano extraídos de una web de compraventa, procedentes tanto de vendedores privados, como de vendedores profesionales.

## Representación gráfica

El siguiente esquema trata de identificar visualmente la estructura del dataset y alguno de los resultados que podrían extraerse de su análisis. A la izquierda de la figura se observa un registro del dataset que identifica el modelo, precio y número de referencia del vehículo, así como otras características relacionadas con el vendedor y datos técnicos del vehículo. Tras un proceso de análisis, se puede extraer conocimiento del mercado de segunda mano, tal y como se representa en la parte derecha del ejemplo.

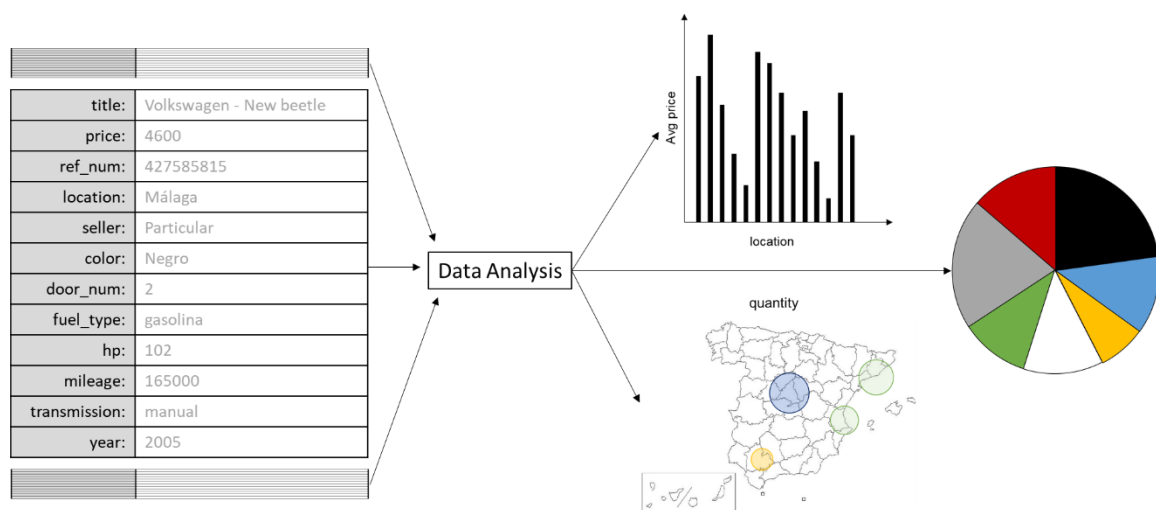


Figura 2. Esquema para ilustrar el dataset. Fuente: Elaboración propia.

## Contenido

Cada registro del dataset incluye los siguientes campos:



- Title: String. Nombre con el que se refiere al coche, generalmente es una combinación de la marca y el modelo unido con un guion. En algunos casos incluye además alguna de las características del modelo como variante y tipo de motor.
- Price: Integer. Precio del vehículo.

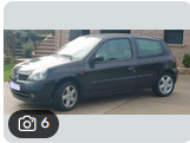
- ref\_num: Integer. Identificador que la web asigna a cada anuncio.
- location: String. Ciudad en la cual se encuentra el vehículo.
- seller: String (Variable Categórica). Puede tomar dos valores: particular o profesional.
- color: String. Color de la carrocería del vehículo.
- door\_num: Integer. Número de puertas del vehículo.
- fuel\_type: String. Combustible que emplea el vehículo. Por lo general, gasolina o diésel.
- hp: Integer. Caballos de potencia que tiene el motor del vehículo.
- mileage: Integer. Kilómetros recorridos por el coche.
- transmisión: String (Variable Categórica). Puede tomar dos valores: manual o automático.
- year: Integer (Date.Year). Año de primera matriculación del vehículo.

De manera habitual, el mercado de segunda mano de vehículos es muy dinámico. La renovación de los anuncios de oferta online evoluciona, cambia y se renueva de manera prácticamente diaria. Así, un dataset extraído en una fecha particular solamente es representativo del estado del mercado actual por un periodo limitado de tiempo, considerando, además, las amplias posibilidades de marcas y modelos de vehículos disponibles en el mercado. Por tanto, el periodo en el que se generan nuevos datos no es fijo, sino que depende de cuándo los usuarios introducen o eliminan los anuncios. En el caso particular del dataset que acompaña a esta práctica, éste fue extraído a las 12:00h, del 4 de noviembre de 2021.

OFERTA

r428953543 DESTACADO

**RENAULT - CLIO**  
 Renault en Pontevedra  
 Mos (Santa Eulalia)



Se vende este Renault Clio 1.2i 16V del año 2002 con 100.000kms.Es de tres puertas y dispone de doble airbag frontal y airbags laterales,dirección...

**Leer más**



99.000 kms   2002   Manual   5 puertas   75 CV


**1.100 €**

Mensaje   Llamar   Compartir   Favorito   Estadísticas   Denunciar

OFERTA

r420811463 DESTACADO

**FIAT - TIPO 1.6 LOUNGE 88KW 120C DIESEL MJET.AUT.SW**  
 Fiat en Pontevedra  
 A guarda/la guardia



Vehículo nacional con historial de mantenimiento , cristales oscuros traseros, sistema de navegación con pantalla táctil , cámara de marcha atrás , control d...

**Leer más**

PROFESIONAL

129.000 kms   2018   Automático   5 puertas   120 CV

Precio al contado

**11.300 €**

IVA Incluido

Precio financiado

**9.800 €**

Mensaje   Llamar   Compartir   Favorito   Estadísticas   Denunciar

Figura 3. Ejemplo de anuncio particular y profesional en la web milanuncios.com.

Para la generación del dataset, se ha hecho uso del lenguaje de programación Python, empleando la librería Scrapy. Con el fin de gestionar elementos que el sitio web carga dinámicamente con JavaScript, adicionalmente se ha hecho uso de la librería Selenium. De esta manera, se han podido recoger todos los datos de los anuncios de forma correcta.

## Agradecimientos

Los datos que forman parte del dataset obtenido provienen de la web de anuncios [www.milanuncios.com](http://www.milanuncios.com).

Los sitios web de compraventa, cuyo modelo de negocio pasa por la publicación de anuncios de terceros, suelen tener contenido aislado y propio. Esto tiene sentido ya que sus ganancias dependen del tráfico que los anuncios generan y, raramente, se mezclan con competidores.

Hemos encontrado diferentes proyectos de scraping relacionados con la temática de nuestra propuesta de proyecto:

1. Web scraping para Machine Learning
  - a. Objetivo: Pagar menos dinero al comprar un coche de segunda mano.
  - b. Lenguaje empleado: Python.
  - c. Librerías usadas: Request y BeautifulSoup.
  - d. Web target: Ejemplo genérico para una web inventada.
  - e. Año: 2020
  - f. Link: <https://todoia.es/web-scraping-para-machine-learning/>
2. Cochista
  - a. Objetivo: Extraer anuncios de coches.net.
  - b. Lenguaje empleado: R.
  - c. Librerías usadas: Tidyverse, Rvest y Httr.
  - d. Web target: [www.coches.net](http://www.coches.net)
  - e. Año: 2018
  - f. Link: <https://github.com/hmeleiro/cochista>
3. Scrapeo coches.net
  - a. Objetivo: Extraer anuncios de coches.net.
  - b. Lenguaje empleado: Python (scraping) y R (visualización).
  - c. Librerías usadas: Request, BeautifulSoup y Tidyverse.
  - d. Web target: [www.coches.net](http://www.coches.net)
  - e. Año: 2018
  - f. Link: <https://wiki.montera34.com/taller-web-scraping-hirakilabs/coches>
4. Milanuncios
  - a. Objetivo: Scraper y autorenovador de anuncios
  - b. Lenguaje empleado: Python.
  - c. Librerías usadas: BeautifulSoup y Selenium.
  - d. Web target: [www.milanuncios.com](http://www.milanuncios.com)
  - e. Año: 2018
  - f. Link: <https://github.com/mondeja/milanuncios>

En cuanto a la realización del proyecto, se han tenido en cuenta los siguientes aspectos éticos y legales:

- El sitio web dispone de un robots.txt que sirve de guía de lo que el administrador permite hacer. Aunque este fichero no implica obligatoriedad, se han considerado las directrices que contiene para determinar cómo acceder a los datos.
- Los anuncios de vehículos disponen de información del vendedor como su nombre o su teléfono. Para proteger la identidad de los usuarios, se ha optado por no extraer dicha información. Únicamente se considera el tipo de vendedor: particular o profesional.

## Inspiración

Este dataset puede aportar información interesante que permita analizar el mercado de coches de segunda mano según diferentes parámetros: precio, marca, modelo, tipo de vendedor, provincia, características del coche, etc.

De cara a un estudio del mercado de vehículos de segunda mano a nivel general, un análisis del dataset puede aportar información que permite responder a las siguientes preguntas:

- Para un mismo tipo de coche, ¿hay provincias preferentes en las que comprar? ¿En términos de precio, de disponibilidad, de garantía profesional, ...?
- Para un mismo tipo de coche, ¿compensa comprárselo a un vendedor profesional? ¿Es más caro o es más barato?
- Si se desea vender un coche a un profesional, ¿qué marcas o modelos tienen éstos como objetivo de venta posteriormente?
- ¿Cómo influye la antigüedad en el precio de un coche?
- Al igual que ha ocurrido en el mercado de vehículos nuevos por el contexto medioambiental y de restricciones, ¿ha cambiado la proporción de vehículos diésel o gasolina que hay a la venta?
- Penetración de coches con sistemas motores de propulsión alternativa (eléctrica, gas, hidrógeno, etc.) en el mercado de vehículos usados.
- ¿Cuál es la devaluación de un modelo particular de vehículo según su edad?

Se comprueba que aunque el enfoque de base es similar al de los proyectos analizados en el apartado anterior, nuestra propuesta va más allá al tratar de responder preguntas más complejas.

## Licencia

La página web de la que se ha obtenido el dataset está sujeta a propiedad intelectual e industrial. En ese sentido, los datos extraídos se consideran propiedad de un tercero y, en consecuencia, se ha elegido un tipo de licencia que restrinja el uso que se les da a los datos.

El dataset se ha publicado bajo licencia [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/). Esta licencia permite copiar, distribuir, transformar y mejorar los datos, pero se debe hacer uso del dataset bajo los siguientes términos:

- Atribución: se debe dar crédito al autor, enlace a la licencia e indicar los posibles cambios efectuados.
- Uso comercial: no está permitido el uso comercial de los datos.
- Compartir: si se transforman o cambian los datos, se deben compartir bajo la misma licencia.

## Código

Link al código: <https://github.com/d-duran/CarAdScraper.git>

## Dataset

Link al dataset: <https://doi.org/10.5281/zenodo.5651148>

<b>Contribuciones</b>	<b>Firma</b>
Investigación previa	DDO, PMP
Redacción de las respuestas	DDO, PMP
Desarrollo del código	DDO, PMP