

# PRÁCTICA 1

Ricardo Colin Pérez  
Universitat Oberta de Catalunya  
E-mail: [rcolinp@uoc.edu](mailto:rcolinp@uoc.edu)

24 de octubre de 2021

## 1.- Contexto: Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. [0.25]


La información se ha recolectado en el contexto de los automóviles. Para ello, se ha seleccionado la página web <https://www.autocasion.com/>, particularmente, el directorio “/coches-ocasión”. Autocasión es un portal centrado en los contenidos y servicios transaccionales para la compra de vehículos nuevos y de ocasión y en la información de actualidad sobre el mundo del motor. Este portal se sitúa año tras año como uno de los portales líderes de automoción en España con toda la experiencia y el conocimiento que aportan sus más de 20 años de trayectoria.

## 2.- Título: Definir un título que sea descriptivo para el dataset. [0.25]

“cars\_dataset” es el nombre que se le ha dado al archivo .csv que contiene el dataset obtenido. Como se puede observar, se ha optado por un título sencillo para describir el conjunto de datos con información sobre coches usados. Como veremos en los siguientes apartados, esta información estará relacionada con los anuncios de los coches publicados en el portal de Autocasión. De este modo, encontraremos datos relacionados con las características de los vehículos y sus precios.

## 3.- Descripción del dataset: Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido. [0.25]



El título “cars\_dataset” se escogió en base a los registros que se recolectan en el conjunto de datos extraído.




**MERCEDES-BENZ Clase GLA 220d 7G-DCT 1**

Provincia: Pontevedra    Matriculación: 2018    29.490 €  
 Combustible: Diésel    Kilómetros: 59.968 km    Con financiación  
 Cambio: Automático    Potencia: 177 cv    [Calcula tu seguro](#)

Valoración del concesionario  
 TALLERES HERMINDO | 3,5/5 | ★★★★★



  [Contactar](#)



**MINI Mini Cabrio Cooper Aut.**

Provincia: Madrid    Matriculación: 2016    20.300 €  
 Combustible: Gasolina    Kilómetros: 102.011 km  
 Cambio: Automático    Potencia: 136 cv

Valoración del concesionario  
 CarNext.com Madrid | 3,5/5 | ★★★★★

  [Contactar](#)

Como se observa en la anterior imagen, la página web de interés cuenta con una serie de anuncios listados de forma consecutiva. Nuestro objetivo ha sido recolectar la información de cada uno de estos anuncios de modo que cada coche sea un registro de nuestro dataset. De esta forma, cada fila incluirá información sobre el nombre del vehículo, la localización, el año de matriculación, y toda aquella información que vemos en cada uno de estos recuadros.




```

<div id="results.html" class="contenido-listado ng-scope" data-ng-controller="Textiln
ksCtrl">
  <!-- item.1 -->
  <article class="anuncio evaluacion" data-ng-mouseenter="overTextLinkPubli($event)"
data-ng-mouseleave="leaveTextLinkPubli($event)">...</article>
  <!-- item.2 -->
  <article class="anuncio evaluacion" data-ng-mouseenter="overTextLinkPubli($event)"
data-ng-mouseleave="leaveTextLinkPubli($event)">...</article>
  <p class="text-publi">Publicidad</p>
  <div class="ng-non-bindable mega-publi" id="boton_native" data-google-query-id="COIU
hfk6_MCFV88paQvIU0PA">...</div>
  <!-- item.4 -->
  <article class="anuncio" data-ng-mouseenter="overTextLinkPubli($event)" data-ng-

```

Mediante la herramienta “Inspeccionar” del navegador, somos capaces de acceder a la estructura anidada de la página web y a partir de aquí somos capaces de identificar las etiquetas que se corresponden a cada uno de estos anuncios de coche.

#### 4.- Representación gráfica: Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido. [0.5]

A continuación, se representa gráficamente una muestra del dataset obtenido. Concretamente, en la siguiente tabla se ilustra un ejemplo de los 5 últimos registros de los datos extraídos de <https://www.autocasion.com/> mediante la función de pandas `df.tail()`:

	Name	Location	Year	Combustible	Kilometers	Transmission	Power [CV]	Price [€]	Financed Price	Dealer	Dealer Rating (over 5)
315	FORD Kuga 1.5 ECOBOOST 150 A-S-S 4X2 TITANIUM	Guipúzcoa	2015	Gasolina	89.000	Manual	150	14.900	Sin financiación	COCAR EXPOCACION Servicio Oficial	3,0
316	FORD Ka 1.20 Titanium	Alicante	2009	Gasolina	145.300	Manual	69	4.900	Sin financiación	AUTOMÓVILES CARRUS	4,0
317	BMW Serie 1 118d	Barcelona	2015	Diésel	75.000	Manual	150	17.000	Con financiación	ÉLITE MATARÓ	0,0
318	OPEL Meriva 1.6CDTI S&S Ecoflex Selective	Valencia	2017	Diésel	40.134	Manual	110	11.490	Con financiación	SYA MOTOR	3,5
319	SEAT León 1.0 EcoTSI S&S Reference 115	Barcelona	2018	Gasolina	35.000	Manual	115	12.950	Con financiación	GESTIONCAR Girona	2,5

Como se puede observar en el dataframe anterior, el conjunto de datos obtenido esta formado por un total de 320 registros y 11 variables. Estos 11 campos se explicarán con más detalle en el apartado anterior. Sin embargo, como podemos ver en este ejemplo, cada entrada recoge información sobre un anuncio de un coche de segunda mano y ocasión en concreto.

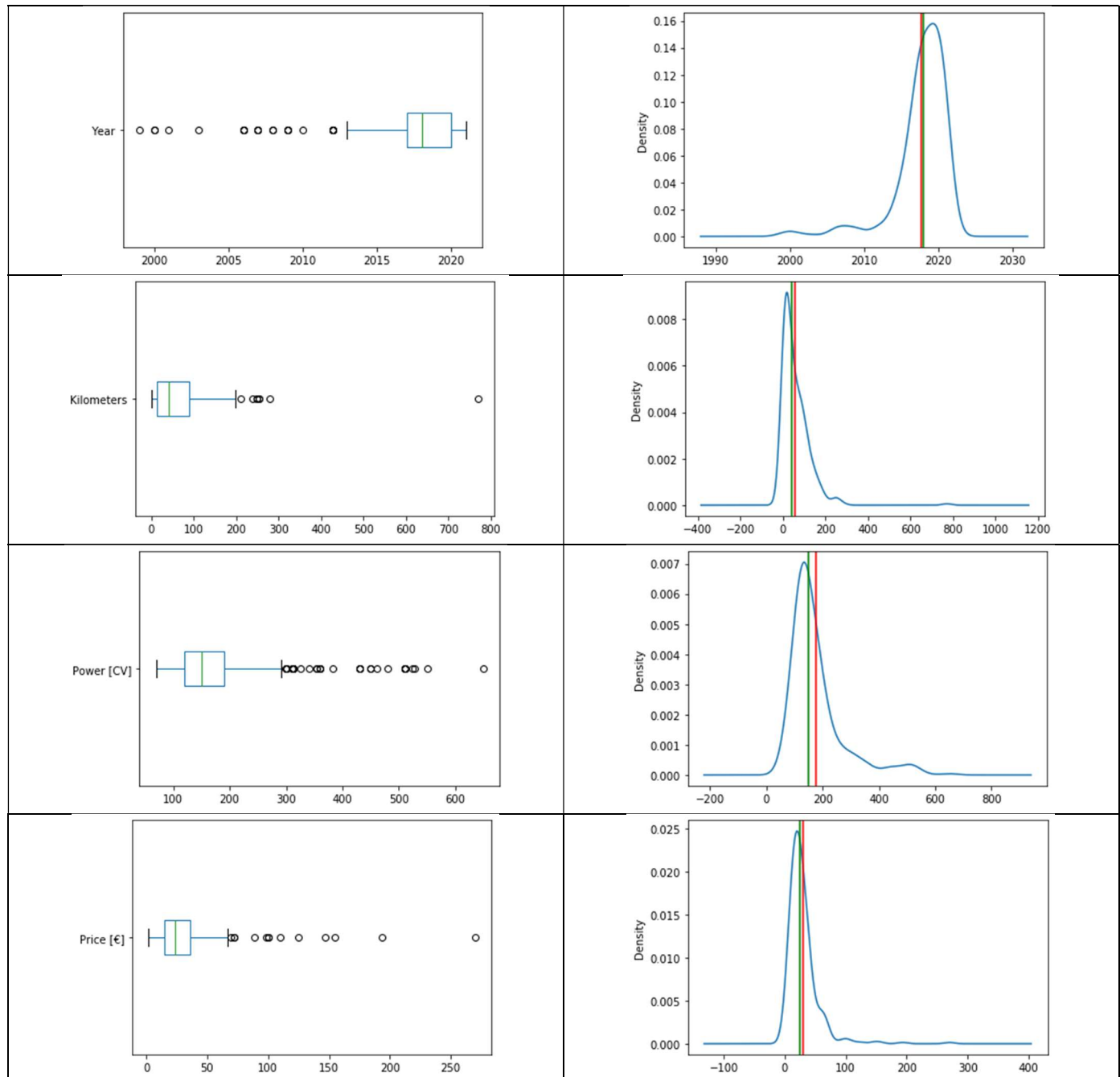
#### 5.- Contenido: Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido. [1]

Como hemos comentado antes, el dataset está formado por 320 observaciones de 11 variables. A continuación, se describe cada uno de estos 11 atributos:

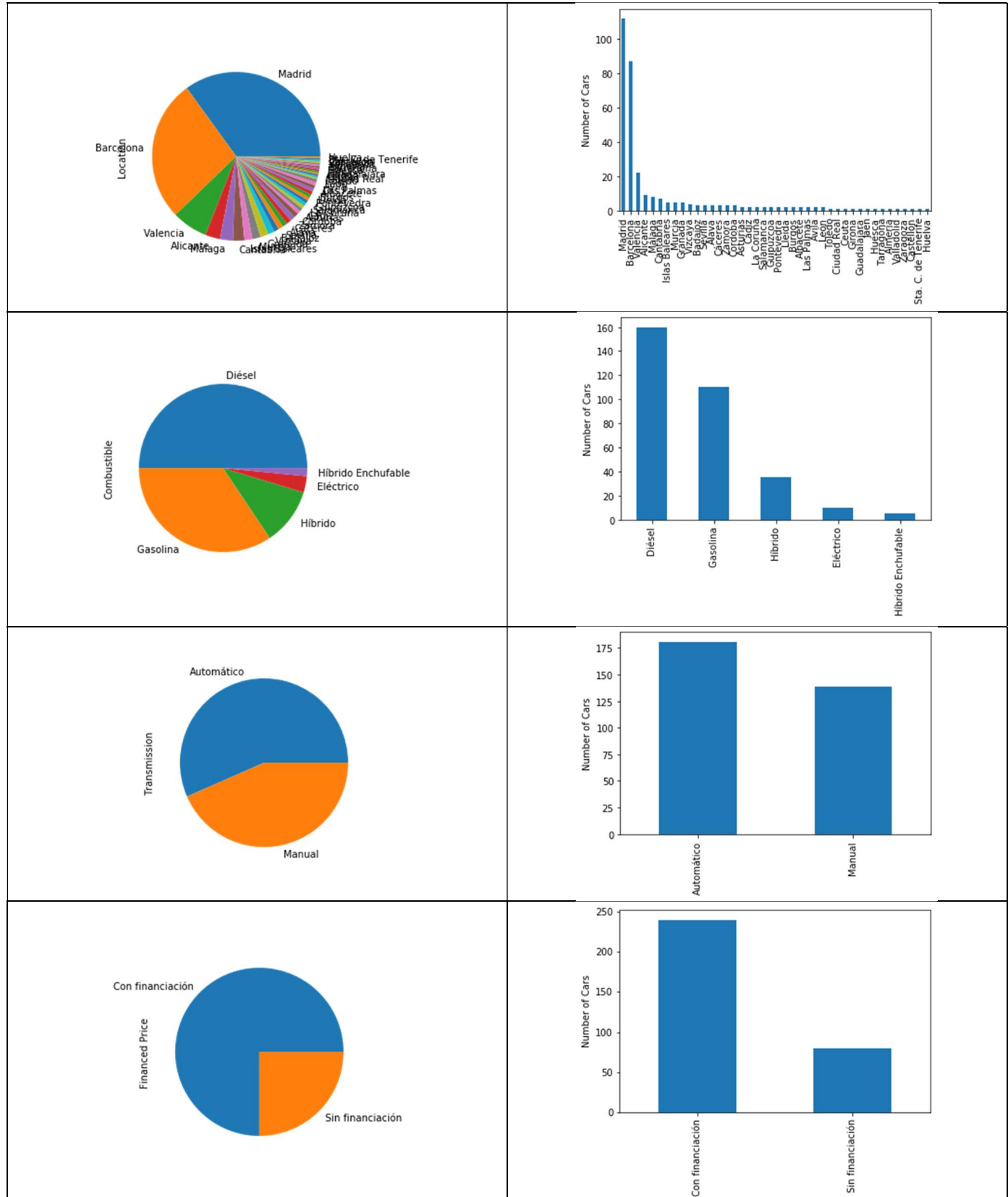
- **Name:** nombre del coche anunciado. Clase: carácter.
- **Location:** ubicación del coche anunciado. Clase: carácter (Madrid, Barcelona, Valencia, Alicante, Málaga, Cantabria, Islas Baleares, Granada, Murcia, Vizcaya, Álava, Badajoz, Cáceres, Zamora, Sevilla, Córdoba, Pontevedra, Salamanca, La Coruña, León, Lleida, Cádiz, Asturias, Guipúzcoa, Las Palmas, Albacete, Ávila, Burgos, Valladolid, Almería, Huelva, Sta. C. de Tenerife, Jaén, Huesca, Castellón, Girona, Toledo, Tarragona, Zaragoza, Guadalajara, Ciudad Real, Ceuta).
- **Year:** año de matriculación del coche anunciado. Clase: entero.
- **Combustible:** tipo de combustible del coche anunciado. Clase: carácter (Diésel, Gasolina, Híbrido, Eléctrico, Híbrido Enchufable).
- **Kilometers:** cantidad de kilómetros del coche anunciado. Clase: float.

- **Transmission:** tipo de transmisión del coche anunciado. Clase: carácter (Automático, Manual).
- **Power [CV]:** cantidad de caballos de potencia del coche anunciado. Clase: entero.
- **Price [€]:** precio del coche anunciado. Clase: float.
- **Financed Price:** financiación, o no, del precio del coche anunciado. Clase: carácter (Con financiación, Sin financiación).
- **Dealer:** concesionario del coche anunciado. Clase: carácter.
- **Dealer Rating (over 5):** puntuación sobre 5 del concesionario del coche anunciado. Clase: float.

Además, para comprender la naturaleza de algunos de los campos incluidos en el dataset, se han realizado unas representaciones que proporcionan mucha información visualmente. Primero, se muestran gráficas de algunas variables numéricas ("Year", "Kilometers", "Power [CV]" y "Price [€]"). A la izquierda vemos los diagramas de caja y a la derecha vemos los diagramas de densidad (media en rojo y mediana en verde) de estas mismas variables:



Seguidamente, se muestran gráficas de algunas variables categóricas (“Location”, “Combustible”, “Transmission” y “Financed Price”). A la izquierda vemos los gráficos circulares y a la derecha vemos los diagramas de barras:



En cuanto al periodo de tiempo de estos datos, no hemos fijado ninguna periodicidad para su recolecta. Simplemente, los hemos captado en el momento de ejecutar el código Python (24/10/2021). Como <https://www.autocasion.com/coches-ocasion> se actualiza constantemente con nuevos anuncios, se podrían recoger los datos con un periodo de tiempo de 1 día dado que el código que hemos generado recoge anuncios de 20 páginas con 16 anuncios por página.

Finalmente, daremos una breve descripción de como se han recogido los campos mencionados anteriormente. En resumen, el proceso se puede dividir en tres pasos:

1. Envío de petición HTTP: cuando queremos acceder a la página web de interés (diseñada en lenguaje HTML) a través del navegador, se realiza una petición HTTP. Hay que destacar que, en el código implementado en esta práctica, se ha añadido la información "User-agent" en la cabecera de la petición HTTP con el fin de no ser bloqueados.
2. Envío de respuesta HTTP: cuando el navegador realiza la petición HTTP, el servidor responde incluyendo cabeceras HTTP junto con un documento HTML que contiene la información que queremos rastrear.
3. Conversión de HTML a estructura anidada: por último, el navegador parsea la página web para construir una estructura anidada.

Aunque el web scraping se puede realizar mediante diferentes lenguajes de programación, estos pasos se han llevado a cabo gracias a las librerías de Python *requests* y *beautiful soup*. Posteriormente, una vez conseguimos la estructura anidada del sitio web, navegamos por ella mediante algunos de los comandos más utilizados como *.find()*, o *.find\_all()*. De esta forma, capturamos cada una de las variables listadas anteriormente, con las que formamos el dataframe mostrado en el apartado 4 para finalmente crear el dataset.

**6.- Agradecimientos:** Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto. [1.5]

En cuanto al propietario del conjunto de datos, lo identificaremos a través del mismo propietario de la página web de la cual hemos obtenido la información. Este proceso, también puede ser interesante cuando, por ejemplo, el propietario sea conocido por bloquear los procesos de *web scraping*. Mediante la función *whois.whois()*, llamamos a <https://www.autocasion.com/> para conocer el propietario, obteniendo el siguiente resultado:

print(whois.whois('https://www.autocasion.com/coches-ocasion'))		
"domain name": { "AUTOCASION.COM", "autocasion.com" }	"expiration date": { "2030-01-09 15:17:23", "2030-01-09 04:00:00" }	"org": "sumauto motor s.l."
"registrar": "Onlinenic Inc"	"name_servers": { "NS1-05.AZURE-DNS.COM", "NS2-05.AZURE-DNS.NET", "NS3-05.AZURE-DNS.ORG", "NS4-05.AZURE-DNS.INFO", "ns1-05.azure-dns.com", "ns2-05.azure-dns.net", "ns3-05.azure-dns.org", "ns4-05.azure-dns.info" }	"address": null
"whois_server": "whois.onlinenic.com"	"status": { "clientDeleteProhibited <a href="https://icann.org/epp#clientDeleteProhibited">https://icann.org/epp#clientDeleteProhibited</a> ", "clientTransferProhibited <a href="https://icann.org/epp#clientTransferProhibited">https://icann.org/epp#clientTransferProhibited</a> " }	"city": null
"referral_url": null	"emails": "abuse@onlinenic.com"	"state": "Madrid"
"updated date": { "2020-10-05 09:09:47", "2020-10-05 05:09:46" }	"dnssec": "unsigned"	"zipcode": null
"creation_date": { "2000-01-09 15:17:23", "2000-01-09 04:00:00" }	"name": null	"country": "ES"

A pesar de no haber encontrado análisis anteriores iguales al que se ha llevado a cabo en esta práctica, sí que existen diferentes ejercicios enfocados a la recopilación de datos de páginas web de concesionarios de coches. En mayor o en menor medida, todos y cada uno de ellos ha proporcionado ideas y ha sido de utilidad como guía a la hora de implementar el código. Algunos de estos ejemplos se encuentran explicados en los siguientes videos:

- Web scraping (<https://www.cars.com/>): <https://www.youtube.com/watch?v=NkFc1nWBG0>.
- Web scraping y creación de un modelo (<https://www.kijijiautos.ca/>): <https://www.youtube.com/watch?v=kTpcjYSFDHI>.

Seguindo la guía de la teoría sobre cómo realizar *web scraping*, uno de los aspectos para tener en cuenta durante la fase previa a la extracción de contenido es el archivo *robots.txt*. En este archivo, podremos ver las restricciones a considerar antes de rastrear la página web de interés. A pesar de que no es ilegal ignorar el archivo *robots.txt*, puesto que las restricciones que incluye son solo una sugerencia y nunca una obligación, no es ético pasar desapercibido su contenido. Simplemente añadiendo “/robots.txt” al final del enlace del sitio web que queremos rastrear, accedemos a su contenido. En nuestro caso, el enlace “<https://www.autocasion.com/robots.txt>” nos muestra el siguiente contenido:

User-agent: proximic Disallow: / User-agent: Exabot/3.0 Disallow: / User-agent: GrapeshotCrawler/2.0 Disallow: / User-agent: trendictionbot0.5.0 Disallow: / User-agent: psbot-page Disallow: / User-agent: psbot/0.1 Disallow: / User-agent: WijuBot/1.0 Disallow: / User-agent: GetIntent Crawler Disallow: / User-agent: psbot-image Disallow: / User-agent: istellabot/t.1 Disallow: / User-agent: SMTBot/1.0 Disallow: / User-agent: DeuSu/5.0.2 Disallow: / User-agent: JikeSpider Disallow: / User-agent: baiduspider Disallow: / User-agent: Baiduspider Disallow: / User-agent: BaiDuSpider Disallow: / User-agent: Baiduspider+ Disallow: / User-agent: BaiDuSpider+ Disallow: / User-agent: baiduspider+ Disallow: / User-agent: Yandex Disallow: / User-agent: WeSEE Disallow: / User-agent: * Disallow: /index2.php User-agent: Cliqzbot Disallow: /	#actualizacion 29/08/2014 User-agent: * Disallow: /*anno_hasta= Disallow: /*anno_desde= Disallow: /*color_exterior= Disallow: /*combo_carroceria= Disallow: /*combustible= Disallow: /*garantia= Disallow: /*kms_desde= Disallow: /*kms_hasta= Disallow: /*margen_anno= Disallow: /*margen_kms= Disallow: /*margen_potencia= Disallow: /*margen_precios= Disallow: /*position= Disallow: /*precio_desde= Disallow: /*precio_hasta= Disallow: /*puertas= Disallow: /*tipovehiculo= Disallow: /*tipovendedor= Disallow: /*carroceria= Disallow: /*texto_modelo= Disallow: /*metalizado= Disallow: /*gestion-alertas.php* Disallow: /*/alertas/* Disallow: /*/4900/vocento.autocasion/* Disallow: /4900/vocento.autocasion/* Disallow: /*/coches-nuevos/ficha_tecnica/* Disallow: /*/coches-nuevos/equipamiento/* Disallow: /actualidad/tag/* Disallow: /*/gestion-alertas.php* Disallow: /actualidad/noticias/*/efe- Disallow: /actualidad/noticias/*/europapress- Disallow: /actualidad/noticias/*/reuters- Disallow: /actualidad/?s= Disallow: /coches-segunda-mano/rolls-royce-1-ocasion Disallow: /coches-segunda-mano/rolls-royce-1-ocasion* Disallow: /coches-segunda-mano/land-rover-1-defender-ocasion/ Disallow: /coches-segunda-mano/land-rover-1-defender-ocasion* Disallow: /coches-km0/land-rover-1-km0* Disallow: /coches-km0/rolls-royce-1-km0* Disallow: /*fecha-desde- Disallow: /*fecha-hasta- Disallow: /*precio-desde- Disallow: /*precio-hasta- Disallow: /*km-desde- Disallow: /*km-hasta- Disallow: /*potencia-desde- Disallow: /*potencia-hasta- Disallow: /profesional/*/* Disallow: /profesional/*sort* Disallow: /politica-cookies Disallow: /politica-privacidad Disallow: /movil* Disallow: /pixel Disallow: /marcas/*-9999999 Disallow: /*?sponsor_id= Disallow: /*?interactive_banner_id= Disallow: /actualidad/search/* Disallow: /actualidad/page/* Allow: /coches-segunda-mano/particular	#actualizacion 24/11/2016 Allow: /actualidad/*combustible* Disallow: /actualidad/page/*combustible* Allow: /actualidad/*garantia* Allow: /actualidad/*puertas* Allow: /actualidad/*carroceria* Allow: /actualidad/*metalizado* Allow: /actualidad/*galeria* Sitemap: https://www.autocasion.com/actualidad/sitemap_index.xml Sitemap: https://www.autocasion.com/uploads/sitemap.xml Disallow: /rate/* Disallow: /rated/* Disallow: /ads/galeria/* Disallow: /*submodelo= Disallow: /particular/* #actualizacion 19/10/2021 Allow: /tasacion-de-coches Allow: /vender-coche Allow: /vender-coche-rapido
--	---	--

Después de revisar cada una de estas líneas, vemos que no hay exclusión del directorio “/coches-ocasión” con el cual queremos trabajar.



**7.- Inspiración:** Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6. [1.25]

Los datos obtenidos durante esta práctica pueden ser utilizados para muchos propósitos. Algunos de los ejemplos más destacados son:

- Para analizar la competencia: como todos sabemos, el mercado de la automoción es muy grande. Son muchas las empresas que comercializan con vehículos a precios muy competitivos. Por ello, es interesante contar con información de las características de los vehículos que venden otras empresas, así como del precio que le asignan.
- Para generar modelos de predicción: Como se ha visto en los anteriores apartados, el dataset obtenido contiene información de muchos vehículos, tanto de sus características como del valor por el que han sido tasados. De cara a la necesidad de generar un modelo capaz de tasar nuevos vehículos, contamos con ejemplos que nos permitirán entrenar un algoritmo con el fin de realizar predicciones de precios. Si tenemos en cuenta que las variables de entrada son las características del coche ("Year", "Combustible", "Kilometers", "Transmission", "Power [CV]"), y que la variable de salida es el precio ("Price [€]"), es relativamente sencillo generar un modelo a través de los diferentes ejemplos recopilados con el fin de tasar el precio de nuevos coches a partir de sus características.

Comparando los análisis similares presentados en el apartado 6, para el primer caso encontramos que simplemente se desea resumir la información contenida en diferentes páginas como hemos hecho en nuestra práctica. Sin embargo, partiendo de que las páginas webs analizadas son diferentes, los campos que se extraen en cada caso son también diferentes. Para el segundo caso, nos encontramos de nuevo con algo similar a lo que hemos realizado en este ejercicio a diferencia de que se seleccionan otros datos para recolectar, y que además posteriormente también se utilizan para generar un modelo de machine learning. En comparación al modelo que nosotros generaríamos, donde utilizaríamos muchas variables para entrenar el algoritmo seleccionado, en el trabajo de <https://www.kijijiautos.ca/>, sólo se emplean las variables del fabricante del vehículo, el modelo y el año para predecir su valor.

**8.- Licencia:** Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección: [1]

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Other (specified above).
- Unknown License.

La licencia escogida para la publicación de este conjunto de datos ha sido CC BY-SA 4.0 License. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado.

Para encontrar estas cláusulas se ha visitado la siguiente dirección: <https://creativecommons.org/licenses/by-sa/4.0/deed.es>. En resumen, esta licencia da libertad de compartir y adaptar el material bajo los términos de atribución adecuados y de distribuir la contribución bajo la misma licencia original.

**9.- Código:** Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. [2]

Repositorio Git con el código con el que se ha generado el dataset: <https://github.com/RicardoCoPe/Pr-ctica-1-Web-scraping>.

**10.- Dataset:** Publicar el dataset obtenido en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI. [2]

Enlace del DOI: 10.5281/zenodo.5608503.

Contribuciones	Firma
Investigación previa	RCP
Redacción de las respuestas	RCP
Desarrollo del código	RCP