


Tipología y ciclo de vida de los datos

---

# PRAC 1:WEB SCRAPING


## PRECIOS DE LOS DERIVADOS DEL PETRÓLEO EN ESPAÑA

Alumno: Javier Jiménez Reyes  
Abril 2019

	Universitat Oberta de Catalunya	Alumno      Javier Jiménez Reyes
	Máster Universitario de Ciencia de Datos.	NIF      45832279H
	Tipología y ciclo de vida de los datos	PRAC 1

## Tabla de contenido

1. Introducción .....	2
2. Contexto .....	2
3. Datos y características.....	2
4. Captura de datos .....	2
5. Dataset output.csv .....	3
6. Resultados .....	4
7. Agradecimientos .....	5
8. Licencia .....	5
9. Recursos .....	5

	Universitat Oberta de Catalunya	Alumno Javier Jiménez Reyes
Máster Universitario de Ciencia de Datos.	NIF	45832279H
Tipología y ciclo de vida de los datos		PRAC 1

## 1. Introducción

Es una realidad el hecho que el precio de los combustibles va variando con el tiempo a medida que el mercado los va regulando. De este modo estos suben y bajan en función de varios condicionantes, pero ¿cuál ha sido su evolución?

El objetivo que se persigue en esta práctica consiste en capturar mediante la técnica del **web scraping**, datos que permitan dar respuesta a la pregunta formulada.

## 2. Contexto

El sitio web [Datosmacro.com](http://Datosmacro.com) cuenta con un amplio repositorio de datos de distinta temática, los cuales actualiza en base a un seguido de [webs externas](#) que utiliza como **fuentes de información primarias**.

Entre todos estos repositorios se encuentra el evolutivo de los [Precios de los derivados del petróleo en España](#). En este se indica la evolución del precio de los distintos combustibles fósiles **Super95**, **Diesel** y **Diesel para calefacción** con y sin impuestos añadidos a su valor, dentro del contexto del consumidor final.

## 3. Datos y características

En el desarrollo de la práctica se realizará la captura de la totalidad de los atributos (**Fecha**, **Super 95**, **Super 95 (Sin imp.)**, **Diesel**, **Diesel (Sin imp.)**, **Diesel Cal.**, **Diesel Cal. (Sin imp.)**), cuyo valor se expresa en €, para todas las fechas de la web en las que constan registros (2005 al 2019).


Posteriormente a su captura, los datos serán agrupados por años indicando por cada uno de los combustibles el valor promedio calculado para ese período.

Tras la agrupación por años se graficarán únicamente **los valores con impuestos** que son los que muestran el valor de caras al consumidor final. Estos datos permitirán conocer el evolutivo del valor de los derivados del petróleo en España y a su vez poder observar cual ha sido su tendencia a lo largo de los últimos años.

## 4. Captura de datos

Para la captura de datos mediante la técnica de **web scraping** se ha utilizado el lenguaje de programación **Python**. Para la ejecución del *script* [Code.py](#), es necesario la instalación del sistema de gestión de paquetes **PIP** para Python, así como las siguientes bibliotecas:

- **pip install requests**: Biblioteca para realizar la petición (target) a la página web y facilita la interacción con esta.
- **pip install BeautifulSoup4**: Biblioteca para realizar análisis de documentos HTML.
- **pip install lxml**: Biblioteca para realizar el procesado del HTML a lxml.

	Universitat Oberta de Catalunya	Alumno	Javier Jiménez Reyes
Máster Universitario de Ciencia de Datos.		NIF	45832279H
Tipología y ciclo de vida de los datos			PRAC 1

- ***pip install DateTime***: Biblioteca para la conversión de la variable string a date.
- ***pip install matplotlib***: Biblioteca para realizar gráficos
- ***pip install pandas***: Biblioteca para la generación de DataFrames (hojas de datos) que permitan exportar a .csv.

Los aspectos que se han tenido en cuenta para la captura de los datos han sido:

1. Las marcas utilizadas:
  - **`\<tbody>`**: Marca donde se encuentra la tabla con los datos.
  - **`\<tr>`**: Marca los distintos registros.
  - **`\<td>`**: Marca las distintas celdas del registro.
2. Se deberá convertir la variable **Fecha** del tipo **string** a tipo **date**.
3. Se deberá eliminar el carácter **€** de los atributos correspondiente al valor de los distintos combustibles.
4. Se deberá convertir el carácter **'** a **.'** de los valores de los distintos combustibles, para posteriormente poder convertir el tipo de variable de **string** a **float**.

Los resultados se almacenan en un archivo **.csv** con el nombre **output.csv**. Indicar que en el código se dejan comentados:

- La opción de graficar los valores de los carburantes sin impuestos.
- Generar un archivo outputII.csv con los registros agrupados por año (son los valores que se utilizan para realizar el gráfico).


## 5. Dataset output.csv

El **dataset** generado se compone de **7 atributos** y un total de **706 registros** estructurados del siguiente modo:

Fecha	Super 95	Super 95 (Sin imp.)	Diesel	Diesel (Sin imp.)	Diesel Cal.	Diesel Cal. (Sin imp.)	Año
-------	----------	---------------------	--------	-------------------	-------------	------------------------	-----

La información que guardan los distintos atributos es:

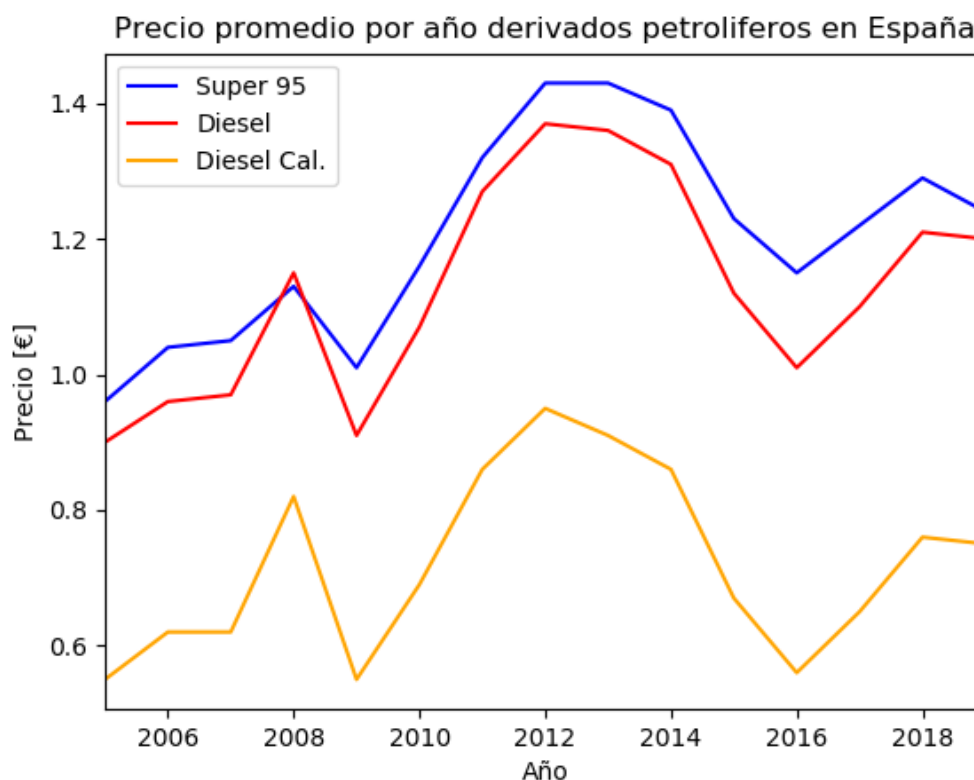
- **Fecha**: Fecha del registro en formato (dd/mm/yyyy). Variable tipo **Date**.
- **Super95**: Coste del combustible Super95. Variable tipo **Float** con dos decimales.
- **Super95 (Sin imp.)**: Coste del combustible Super95 sin impuestos. Variable tipo **Float** con dos decimales.
- **Diesel**: Coste del combustible Diesel. Variable tipo **Float** con dos decimales.
- **Diesel (Sin imp.)**: Coste del combustible Diesel sin impuestos. Variable tipo **Float** con dos decimales.
- **Diesel Cal.**: Coste del combustible Diesel calefacción. Variable tipo **Float** con dos decimales.

 Universitat Oberta de Catalunya	Alumno Javier Jiménez Reyes
Máster Universitario de Ciencia de Datos.	NIF 45832279H
Tipología y ciclo de vida de los datos	PRAC 1

- **Diesel Cal. (Sin imp.):** Coste del combustible Diesel calefacción sin impuestos. Variable tipo **Float** con dos decimales.
- **Año:** Año extraído del atributo **Fecha**.


## 6. Resultados

El siguiente gráfico representa para los años **2005 a 2019** la evolución del valor (con impuestos) promedio por año, de los distintos derivados del petróleo.



Algunos de los aspectos que se pueden comentar de la representación gráfica de los valores son:

- Se aprecia una tendencia ascendente general para cada uno de los combustibles a lo largo de los años.
- La tónica general que se mantiene a lo largo de los años es que el valor del Super 95 es superior al Diesel.
- Se aprecia un cambio de tendencia en la tónica general en el 2008, año en que el valor promedio del Diesel es superior al del Super95. Esta punta corresponde a la subida del 4 de agosto de dicho año en el que el Diesel superó por primera vez el precio de la gasolina.
- Los máximos históricos se dan en los periodos del 2012 al 2014.

 Universitat Oberta de Catalunya	Alumno	Javier Jiménez Reyes
	NIF	45832279H
Máster Universitario de Ciencia de Datos.		
Tipología y ciclo de vida de los datos		PRAC 1

## 7. Agradecimientos

Agradecer al sitio web [Datosmacro.com](https://datosmacro.com) así como el directorio de [webs externas](#) que utiliza para mantener los datos actualizados, el haber facilitado dicha información para la realización de la práctica.

## 8. Licencia

El código y los datos se publicarían bajo una licencia CC BY-NC-SA 4.0, es decir, licencia **Creative Commons Non-Commercial y Share-Alike**. Bajo esta licencia tenemos las siguientes condiciones:

1. El material se puede copiar, redistribuir, adaptar y modificar en cualquier medio o formato, fomentando la colaboración y la filosofía open-source.
2. Se incluye el término de la atribución, para que al replicar o usar el contenido se deba citar a la fuente original.
3. El término Share Alike de la licencia nos garantiza que cualquier modificación o uso de este material se publique bajo la misma licencia, en aras de promover la colaboración.
4. El término Non Commercial determina que el material no puede ser usado para fines comerciales y por tanto su ámbito se reduce puramente al académico.

## 9. Recursos

Los recursos utilizados para el desarrollo de la práctica han sido:

- *Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.*
- *Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.*