

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. Contexto	2
2. Título del dataset.....	2
3. Descripción del dataset	3
4. Representación gráfica.....	4
5. Contenido	5
6. Agradecimientos	6
7. Inspiración	6
8. Licencia	6
9. Código.....	7
10. Dataset	7

	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. Contexto

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.

El sitio web [Datosmacro.com](https://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.

El contenido en la web se encuentra separado por años, de forma que para la consulta de los distintos registros se debería ir consultando una por una las páginas del repositorio e ir copiando estos datos a una hoja de trabajo externa (ejemplo: arhico excel). Se trata de un proceso operativamente lento, al que haría falta añadir el tiempo que se deberá invertir a posteriori para adecuar los valores copiados para poder extraer información de ellos.

En base al contexto en que se encuentran los datos necesarios para dar respuesta a la pregunta formulada, se puede observar que es un caso adecuado para aplicar la técnica de **web scraping**.

2. Título del dataset

Como título para el *dataset* se ha escogido: **EVOLUCIÓN DE PRECIOS DE LOS DERIVADOS DEL PETRÓLEO EN ESPAÑA**. El motivo de escoger este título es porqué describe el objetivo principal planteado: conocer **cómo han evolucionado los precios de los combustibles derivados del petróleo en España**.

 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

3. Descripción del dataset

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).


El *dataset* finalmente generado tras la captura, se compone de **7 atributos** y un total de **705 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.
- **Diesel Cal. (Sin imp.):** Coste del combustible Diesel calefacción sin impuestos. Variable tipo **Float** con dos decimales.
- **Año:** Año extraído del atributo **Fecha**. Campo calculado.

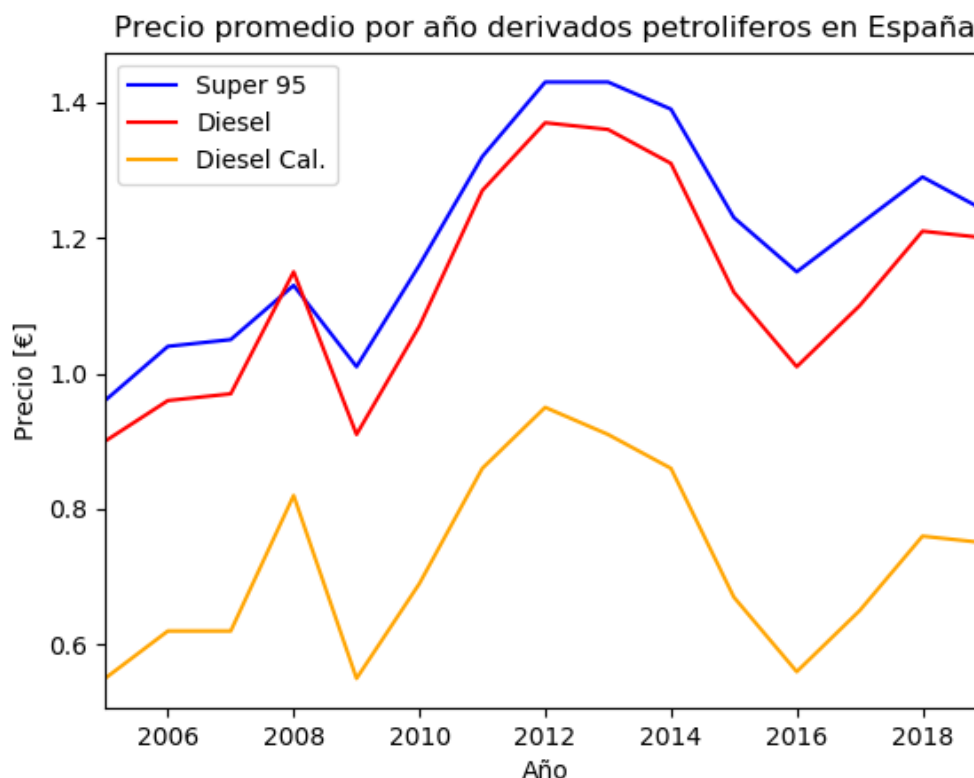
Posteriormente a su captura los datos serán **agrupados por años**, indicando por cada uno de los combustibles el **valor promedio** calculado para ese período (**ejemplo:** para el 2005 se recogerá el valor promedio, por combustible, de los precios que ha tenido en ese año). Estos datos permitirán visualizar la evolución 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.

	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

4. Representación gráfica


Tras la agrupación por años mencionada en el apartado anterior, se graficarán únicamente **los valores con impuestos** que son los que muestran el valor de caras al consumidor final.

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

5. Contenido


Los campos que forman el *dataset* son los descritos en el apartado de **Descripción del dataset** y que corresponden a los siguientes 8 atributos:

- **Fecha.**
- **Super95.**
- **Super95 (Sin imp.).**
- **Diesel.**
- **Diesel (Sin imp.).**
- **Diesel Cal..**
- **Diesel Cal. (Sin imp.).**
- **Año** (campo calculado).

El periodo de datos, también expuesto en el apartado de **Descripción del dataset**, va desde el año **2005** al año **2019**(actualidad).

Estos datos han sido capturados mediante la técnica de **web scraping** utilizando como lenguaje de programación **Python**. Los pasos seguidos para la captura de los datos han sido:

1. Carga de los paquetes y herramientas necesarias para la ejecución del código (especificadas en el apartado **Código**)
2. Inicialización de las listas/variables donde se almacenarán los datos capturados.
3. Lectura del robots.txt de la web [Datosmacro.com](https://datosmacro.com) e impresión para su lectura.
4. Modificación del *user-agent* mediante una función que itera el usuario para evitar bloqueos por *default*.
5. Se genera un bucle mediante tres funciones **for** para la captura de todos los datos:
 - Primer **for**: recorre las distintas páginas que representan los distintos años de capturas.
 - Segundo **for**: recorre los distintos registros de la tabla existente en la página seleccionada por el primer **for**.
 - Tercer **for**: recorre, dentro del registro seleccionado en el segundo **for**, cada una de las celdas para capturar el valor que contienen.
6. Se crea el *dataset* donde se almacenen los datos almacenados en las distintas listas generadas en el paso **2**.
7. Se crea un nuevo campo con el nombre **Año** a partir del campo **Fecha** del *dataset*.
8. Se guarda el *dataset* en un archivo **.csv** con el nombre **output.csv**.
9. Se agrupan los datos en función del campo **Año** creado en el apartado **7** y como valor se coge el promedio de los valores registrados para ese periodo.
10. Se genera y almacena el gráfico con los valores agrupados para los carburantes con impuestos.

	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	

6. 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.

7. Inspiración


El motivo de escoger esta temática es debido a que es un tema que afecta a todo el mundo, ya que todos en mayor o menor medida somos consumidores de este recurso y que debido al constante cambio de su valor, es complejo tener una visibilidad de como a ido evolucionando lo que acaba derivando en únicamente una apreciación de si sube o bajo respecto al valor anterior, por lo que se pierde la visual de que tendencia están teniendo o si el diferencial entre las variaciones de valor negativas y positivas tienden en un computo general hacia arriba o hacia abajo. El hecho de disponer de este histórico de datos permite apreciar estas variaciones a distintos niveles:

- **A corto plazo:** Debido a que se tiene más de una medida dentro del mismo mes en el mismo año, se puede apreciar no tan sólo el número de veces que el valor de estos carburantes se han movido, si no en qué medida (que es la sensación de movimiento que la persona de a pie es capaz de apreciar).
- **A medio plazo:** Al tener mediciones por cada uno de los meses del año se puede apreciar cual ha sido la tónica general a lo largo del año así como apreciar en qué momento se han generado subidas (temporadas de vacaciones en el que la gente coge más el coche, invierno en el que se usa la calefacción,...).
- **A largo plazo:** Es la **visual utilizada para responder al objetivo planteado en la práctica**, conocer cómo ha evolucionado el valor de los carburantes, cual está siendo su tendencia (está creciendo constantemente, uno crece más que el otro,...) o por ejemplo verificar si existen cambios de tendencia (¿ha sido siempre más cara la gasolina que el diesel?).

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.

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

9. Código

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:

Librería	Funcionalidad
pip install requests	Paquete para realizar la petición (target) a la página web y facilita la interacción con esta.
pip install BeautifulSoup4	Paquete para realizar análisis de documentos HTML.
pip install lxml	Paquete para realizar el procesado del HTML a lxml.
pip install DateTime	Paquete para la conversión de la variable string a date.
pip install matplotlib	Paquete para realizar gráficos
pip install pandas	Paquete para la generación de DataFrames (hojas de datos) que permitan exportar a .csv.
pip install urllib3	Paquete para la detección de errores en la carga de la url.
pip install syspath	Paquete para establecer fácilmente rutas comunes y no tener que realizar muchas manipulaciones de ruta.
pip install fake-useragent	Paquete para aleatorizar el user-agent y evitar bloqueos por default.

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**.

10. Dataset

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).