

Práctica 1

TOPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

JUAN FRANCISCO NIETO MENDOZA

MARTA GÓMEZ GALÁN

Tabla de contenido

Introducción	2
Pregunta 1 – Contexto	2
Pregunta 2 – Título	3
Pregunta 3 – Descripción del dataset	3
Pregunta 4 – Representación gráfica	4
Pregunta 5 – Contenido	5
Pregunta 6 – Agradecimientos	6
.....	6
Pregunta 7 – Inspiración	7
Pregunta 8 – Licencia	8
Pregunta 9 – Código	9
Pregunta 10 – Dataset	10
Tabla de Contribuciones	10

Introducción

Tras la lectura de los textos propuestos para la asimilación de contenidos del módulo y la lectura del enunciado de la Práctica 1 y sus correspondientes enlaces de información adicional se nos propone una serie de cuestiones que evaluarán nuestro desempeño con la asignatura y nuestra capacidad analítica y sintética.

A continuación, con objeto de responder a las preguntas lo más directamente posible se enumerarán las distintas preguntas y se responderán a estas de manera secuencial tal como aparecen en el enunciado de la PEC.

Pregunta 1 – Contexto

Explicar en qué contexto se ha recolectado la información.
Explicar por qué el sitio web elegido proporciona dicha información.

El conjunto de datos aquí generado consta de los precios de los derivados del petróleo de los últimos años de los países miembros de la Unión Europea y Reino Unido.

Dicho conjunto de datos se ha recolectado del sitio web <https://datosmacro.expansion.com/>, perteneciente al grupo Expansión (<https://www.expansion.com/>), canal de noticias especializado en economía e información de mercados. A través de dicho portal, Expansión pone a disposición del usuario datos de diversas variables económicas y socio demográficas previamente recolectados de los organismos oficiales de los distintos países y zonas a las que corresponden los datos (<https://datosmacro.expansion.com/legal/acerca-de>).

La elección del Sitio Web se basa en la ventaja de que, a pesar de que la información es una mezcla de varios conjuntos de datos obtenidos de portales públicos, esta ha sido unificada en cuanto a formato y es accesible mediante un único sitio web. De esta manera recolectar la información de todos los países de la Unión Europea y Reino Unido es una labor mucho más asequible para nosotros, como científicos de datos, que de otra manera supondría navegar a través de los distintos organismos oficiales de los distintos países incluidos en el análisis. Y si quisiéramos automatizarlo tendríamos que crear muchos “scrapers” en lugar de uno.

Pregunta 2 – Título

Definir un título que sea descriptivo para el dataset.

Título descriptivo del proyecto:

“Evolución de los precios de los hidrocarburos en Europa y Reino Unido desde Enero 2005 hasta la actualidad”.

Pregunta 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.

En el dataset se incluyen los precios (en euros), con y sin impuestos, de los hidrocarburos Super 95, Diesel y Diesel Calefacción de los actuales países miembros de la Unión Europea y Reino Unido, ordenados por país y fecha. En el dataset se refleja la evolución de los precios por semana desde enero 2005 a la actualidad.

Pregunta 4 – Representación gráfica

Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

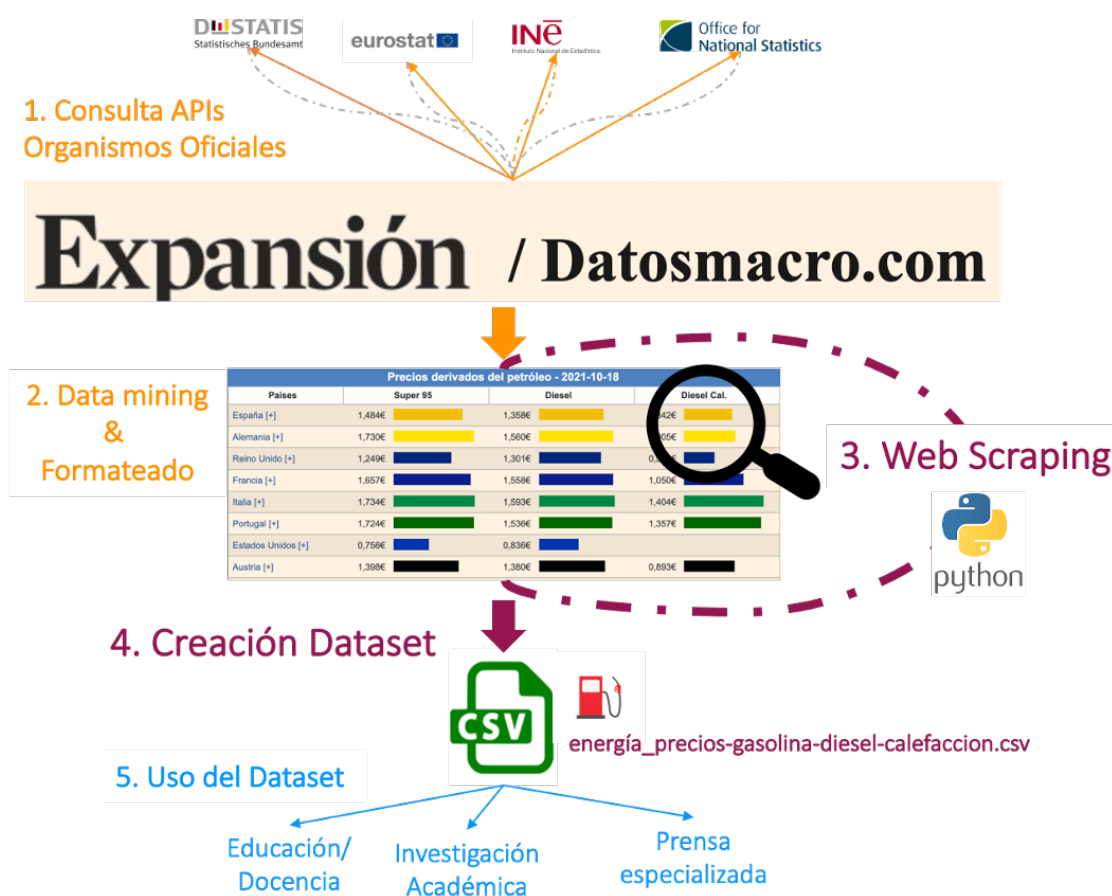


Figura 1. Esquema descriptivo de las distintas etapas para la creación del dataset incluido en el proyecto “Evolución de los precios de los hidrocarburos en Europa y Reino Unido desde enero 2005 hasta la actualidad”. En naranja se muestra la parte llevada a cabo por el equipo del sitio web Datosmacro.com. En granate se indica la parte del web scraping y la generación llevada a cabo por nuestro equipo. En azul se proponen potenciales entidades que se podrían beneficiar del dataset generado.

Pregunta 5 – Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

El dataset incluye los siguientes campos*:

- *Países*: País origen de los datos;
- *Fecha*: en formato Día/mes/año;
- *Super 95*: precio de la gasolina Super 95 con impuestos;
- *Super 95 (sin Imp.)*: precio de la gasolina Super 95 sin impuestos;
- *Diesel*: precio del gasóleo Diesel con impuestos;
- *Diesel (Sin Imp.)*: precio del gasóleo Diesel sin impuestos;
- *Diesel Cal.*: precio del gasóleo Diesel calefacción con impuestos;
- *Diesel Cal. (Sin Imp.)*: precio del gasóleo Diesel calefacción sin impuestos

* los precios se muestran en euros

Con el fin de obtener un dataset lo más completo posible se incluyen todos los registros incluidos en la web de origen. De acuerdo con lo expuesto en dicha web, los datos originales han sido recolectados de distintas fuentes oficiales, tales como los Institutos Nacionales de Estadística de los distintos países europeos incluidos en el dataset así como la Oficina Europea de Estadística (Eurostat), entre otros (<https://datosmacro.expansion.com/legal/fuentes>). Datosmacros.com incluye los precios por semana y año hasta la fecha actual.

En nuestro caso el dataset se ha creado a partir de los datos almacenados en el Sitio Web mediante un scraper creado para tal propósito. Para el scraper se han utilizado las librerías de Python BeautifulSoup para la manipulación del DOM, Requests para las hacer las peticiones al sitio web y Shutil para almacenar las imágenes obtenidas del sitio web.

Pregunta 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.

Los datos han sido recolectados desde el sitio

<https://datosmacro.expansion.com/energia/precios-gasolina-diesel-calefaccion>.

Como ya se ha explicado en el apartado 1, el propietario del conjunto de datos es el grupo Expansión.

La evolución de los precios de los derivados del petróleo ha suscitado siempre gran interés. De esta manera se pueden encontrar en la literatura un gran número de estudios en los que se aborda dicho análisis desde diversos puntos de vista.

Tales estudios suelen publicarse en la sección de economía de los principales medios de comunicación, incluido Expansión, y suelen ser llevados a cabo por empresas relacionadas con el sector automovilístico o plataformas que ofrecen datos obtenidos de distintas fuentes estatales, tal como la aquí utilizada (Datosmacro.com).

A continuación, se citan alguno de ellos:

1. Cuánto ha bajado el precio de la gasolina en cada país de Europa y cómo está en España
https://www.autopista.es/noticias-motor/cuanto-ha-bajado-precio-gasolina-en-cada-pais-europa-como-esta-en-espana_160378_102.html
2. El precio del diésel y la gasolina sigue subiendo: tendencia y las gasolineras más baratas
https://www.autopista.es/noticias-motor/el-precio-del-diesel-y-la-gasolina-sigue-subiendo-tendencia-y-las-gasolineras-mas-baratas_159272_102.html
3. Evolución del precio del diésel y la gasolina, ¿qué factores influyen?
<https://www.race.es/evolucion-precio-diesel-gasolina>
4. España es el país europeo donde más suben los carburantes
<https://www.motorpasion.com/industria/espana-segundo-pais-europa-donde-ha-subido-gasolina>
5. Así ha evolucionado el precio de la gasolina y el diésel desde el año 2000
<https://www.caranddriver.com/es/coches/planeta-motor/a55433/evolucion-precio-gasolina-y-diesel/>

Este tipo de datos también han servido para la realización de estudios académicos. En el siguiente ejemplo se explora el impacto del precio de la gasolina en el número de accidentes de tráfico (<https://pubmed.ncbi.nlm.nih.gov/31838324/>).

Previo a la extracción de los datos que conforman el dataset aquí adjuntado, se ha procedido a la exploración de los términos legales del sitio web (<https://datosmacro.expansion.com/legal/terminos>).

De acuerdo con lo establecido por Datosmacro.com, el acceso y navegación del Sitio Web es libre siempre y cuando se acepten las condiciones de uso. Además, se ha explorado el archivo robots.text (<https://datosmacro.expansion.com/robots.txt>) comprobándose que no incurrimos en ninguna denegación (disallowance).

Pregunta 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.

El conjunto de datos aquí mostrado permite la realización de un amplio abanico de estudios con un distinto enfoque.

De esta manera se pueden resolver preguntas directas tales como cuál es la evolución del precio de la gasolina en los países europeos en los últimos años (estudio 1 y 5, apartado 6), cual es el país europeo con mayores subidas (estudio 4, apartado 6), etc. Además, se pueden combinar con otras bases de datos para dar respuesta a preguntas más complejas como los factores que pueden influir en los precios (estudio 3, apartado 6) o el impacto del precio de la gasolina en los accidentes de tráfico (<https://pubmed.ncbi.nlm.nih.gov/31838324/>).

Aquí se proponen otras preguntas como:

1. Contrastar la evolución de los precios entre distintos países europeos.
2. Explorar el efecto del Brexit en el precio de los hidrocarburos en el Reino Unido.
3. Estudiar el impacto de los impuestos en el precio de los hidrocarburos e identificar aquellos países con mayor carga de impuestos.
4. Estudiar el impacto del precio de los hidrocarburos en el uso del coche/tráfico.
5. Estudiar una posible interacción entre el precio de la electricidad y el de los hidrocarburos
6. El impacto de los precios de los hidrocarburos y las relaciones internacionales entre países vendedores y compradores de petróleo.

Pregunta 8 – Licencia

Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

A continuación, se muestran los tipos de licencias y sus características en formato de tabla y figura, señalando algunas de las presentes en el enunciado de la práctica.

Tabla 1. Se indican las acciones permitidas según los tipos de licencias. Adaptado de Wikipedia https://en.wikipedia.org/wiki/Creative_Commons_license






License name	Abbreviation	Icon	Attribution required	Allows remix culture	Allows commercial use	Allows Free Cultural Works	Meets the OKF 'Open Definition'
"No Rights Reserved" →	CC0		No	Yes	Yes	Yes	Yes
Attribution	BY		Yes	Yes	Yes	Yes	Yes
Attribution-ShareAlike →	BY-SA		Yes	Yes	Yes	Yes	Yes
Attribution-NonCommercial	BY-NC		Yes	Yes	No	No	No
Attribution-NonCommercial-ShareAlike →	BY-NC-SA		Yes	Yes	No	No	No



Figura 2. Open DataBase Licence (ODbL). You are free To Share, To Create, To Adapt as long as you: Attribute, Share-Alike, Keep open. Se muestra de manera esquemática los principios que rigen dicha licencia. Adaptado de Wikipedia https://en.wikipedia.org/wiki/Open_Database_License

Para este tipo de proyecto se elige la licencia ODbL por la cual se permite la libre distribución del dataset así como su modificación y adaptación siempre y cuando se atribuyan los contenidos del dataset y se publiquen de igual manera los contenidos resultantes del uso del dataset, manteniéndolos de acceso público. De esta manera el dataset original se puede ir mejorando a la vez que se facilita su uso y se crea una comunidad de respeto por las fuentes.

Pregunta 9 – Código

Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

A continuación, se adjunta el enlace al repositorio con el código Python creado para la obtención del dataset así como una breve explicación de los principales puntos de dificultad y los mecanismos llevados a cabo para evitar un posible bloqueo por parte del servidor.

<https://github.com/juaniemen-uoc/scrapper-expansion/tree/main/src>

Como principales puntos de dificultad podemos destacar 1) la navegación autónoma del algoritmo y su carácter genérico al realizarse con un algoritmo recursivo final y 2) la inclusión de un scraper de imágenes adicional que incluye las imágenes de las banderas de los países incluidos en el dataset así como la manipulación del DOM.

Para evitar un posible bloqueo del web scraper generado hemos llevado a cabo diversas técnicas:

1. En el caso de web scraper principal (general_scraper) se requieren un total de 514 llamadas para generar el csv completo. Dado que durante el diseño del código se han tenido que realizar muchas pruebas se ha incluido en el código “requests_cache”, en el que los datos se almacenan durante 24h evitando así tener que conectarse al servidor durante las sesiones de prueba de código y levantar sospechas por parte del servidor y, evitando, por tanto, un posible bloqueo durante esta fase.
2. Para evitar saturar el servidor y un posible bloqueo hemos implementado un tiempo aleatorio entre las distintas peticiones. Para ello hemos utilizado las librerías “random” y “time”.

Pregunta 10 – Dataset

Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

Enlace del DOI publicado en Zenodo donde se puede encontrar el CSV generado tras la práctica:

<https://doi.org/10.5281/zenodo.5647903>

Tabla de Contribuciones

Contribuciones	Firma
Investigación previa	MGG; JFNM
Redacción de las respuestas	MGG; JFNM
Desarrollo del código	MGG; JFNM