doc-pract-1

April 9, 2022

1 Tipología y ciclo de vida de los datos: PRÁCTICA I

Componentes de la práctica:

Jonás Medina Brito (jmedinabrit@uoc.edu)
Miguel Rafael Esteban Martín (mestebanmart@uoc.edu)

1.1 Indice de contenidos

1.2 Contexto

La información se ha recolectado de la página oficial de estadísitica de la Comunidad Europea, Eurostat: https://ec.europa.eu/eurostat/web/main/data/database

Se ha obtado por esta fuente de datos por su validez, fiabilidad y diversidad de acceso a la extracción de datos.

1.3 Título

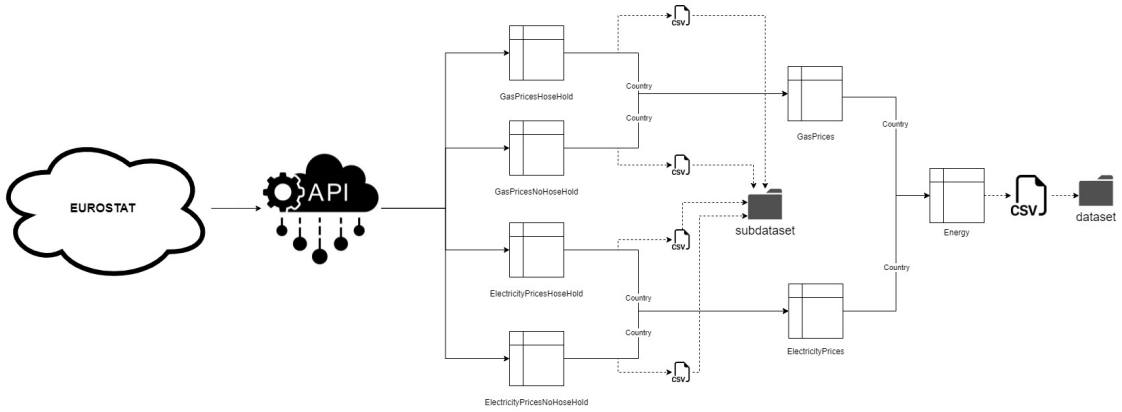
Histórico de precios sobre Energía (Gas y Electricidad) para mercados Familiar y Profesional de la Zona Euro

---

1.4 Descripción del dataset

Histórico de precios sobre Energía (Gas y Electricidad) para mercados Familiar y Profesional de la Zona Euro. El objetivo es obtener el historico de los valores del precio energia, diferenciados entre Gas y electricidad para los años entre el 2017 y el 2021, de tal manera que un registro del dataset, identificado por el país, contenga toda esta información.

1.5 Representación gráfica



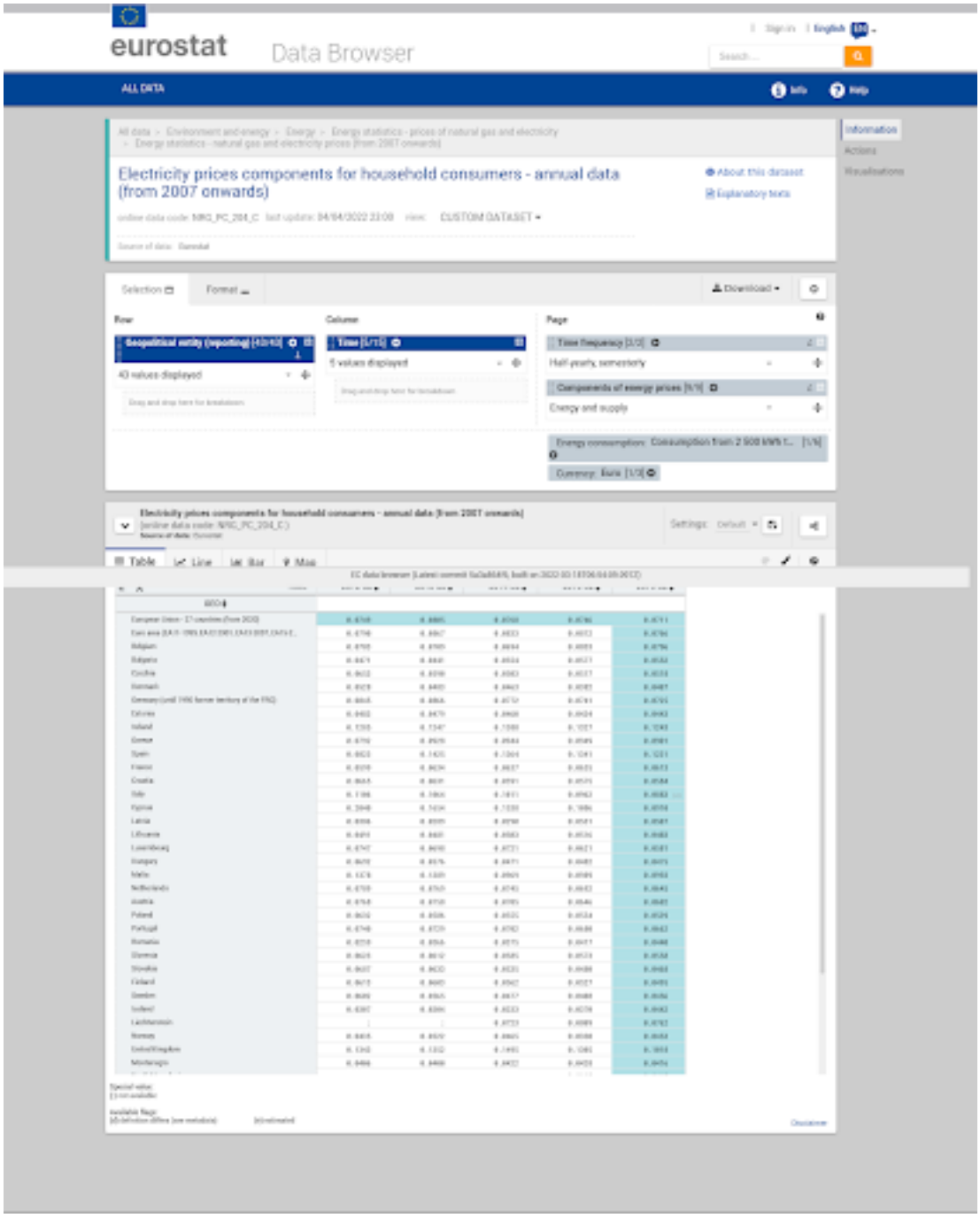
1.6 Contenido

1.6.1 Recogida de datos

Aunque Eurostat tiene una API pública (**1**) que permite obtener los diferentes dataset, en esta práctica hemos decido, obviar esta posibilidad y para acercarnos más al objetivo de la práctica, hacer web scraping de los datos que se ofrecen en su "*Data browser*" accesible desde un navegador.

(1) Eurostat web services: `https://ec.europa.eu/eurostat/en/web/main/data/web-services` (Última vista abril del 2020)

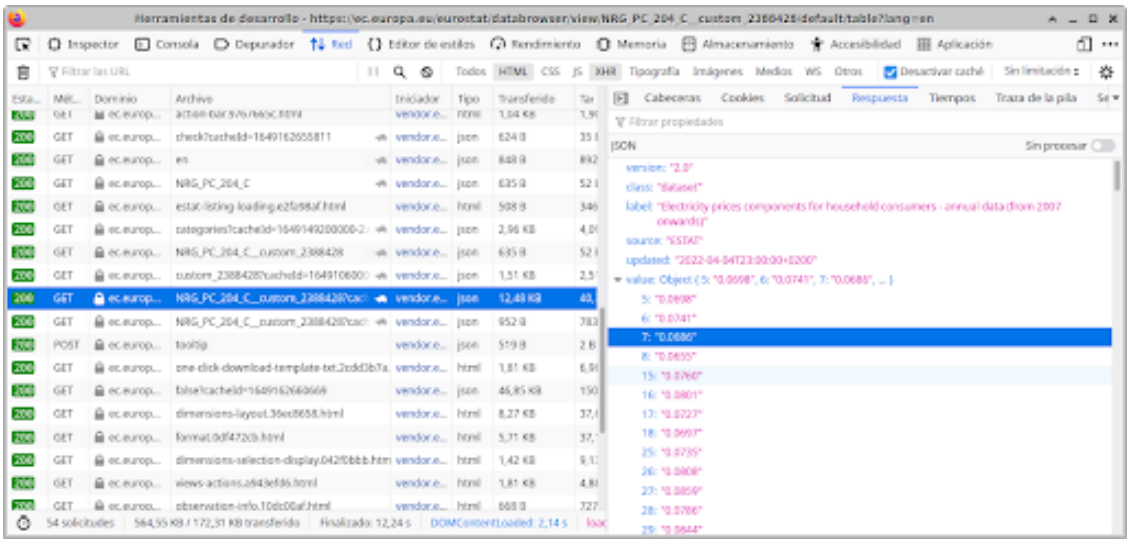
Para explicar el estudio que se ha hecho para la obtención de datos se tomará como ejemplo el dataset de "*Componentes del precio de la electricidad para consumidores domésticos*" (Electricity prices components for household consumers - annual data (from 2007 onwards)



Captura de pantalla de la interfaz "Data Browser" del informe de los componentes del precio de la electricidad para consumidores domésticos

El sitio web de Eurostat utiliza HTML dinámico para generación de la vista, esto quiere decir que, el navegador procesa la información y la presenta generando código HTML y no sólo se limita a representar el código HTML que responde el servidor. Este componente de generación de código por parte del navegador hace complicado el web scraping tradicional y en la mayoría de los casos hace falta la integración con un navegador como hacen librerías como Selenium.

Utilizando ingeniería inversa y las herramientas de desarrollo del navegador (Firefox), se puede sabe cuál es la llamada que devuelve los datos que se representan en la tabla dinámica. Un ejemplo es la petición utilizada en la construcción de la tabla con los datos que se muestra en la página web.



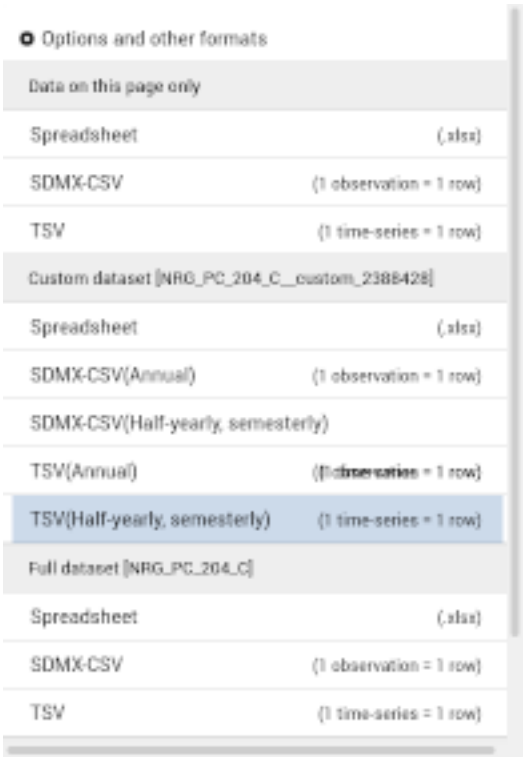
Captura de pantalla de la consola del navegador con la petición de datos de dataset y respuesta JSON

Si se exporta la petición esta es la llamada que se realiza (comando CURL):

```
$ curl 'https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/
LIVE/false/json/en/NRG_PC_204_C__custom_2388428?cacheId=
1649106000000-2.6.3%2520-%25202022-03-17%252005%253A43' …
```

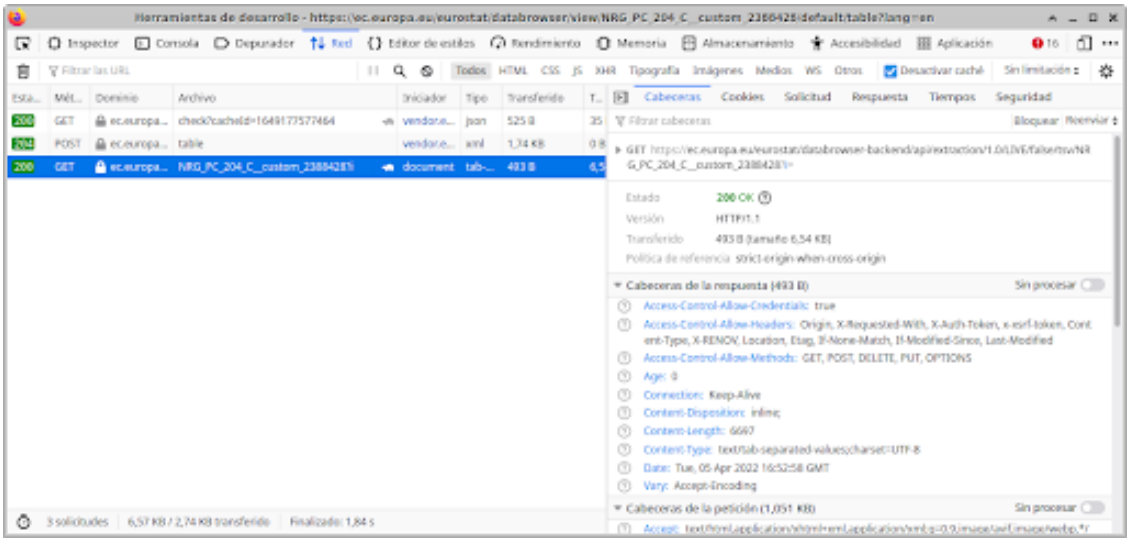
La respuesta de esta llamada es en formato JSON con una estructura específica, difícil de procesar y aprovechar para neustro propósito de acceder a los datos.

Sin embargo la página tiene un formulario de descarga que permite descargar los dataset, en formato TSV mediante el navegador.



Captura de pantalla del enlace de descarga del dataset

Si se hace click en ese enlace la petición que hace el navegador es la siguiente:



Captura de pantalla de la interfaz "Data Browser" de la descarga del informe de los componentes del precio de la electricidad para consumidores domésticos en formato TSV

Qué si se exporta la petición en formato CURL es:

```
$ curl 'https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/LIVE/
```

```
false/tsv/NRG_PC_204_C__custom_2388428?i' …
```

Y que si se ejecuta desde una consola se obtiene los datos en el formato TSV

```
freq,nrg_cons,nrg_prc,currency,geo\TIME_PERIOD  2012-S2     2013-S2     2014-S2     2015-S2
A,KWH2500-4999,NETC,EUR,AL  :    :    :    :    :   0.0000  0.0000  0.0000  :   0.0000
A,KWH2500-4999,NETC,EUR,AT  :    :    :    :    :   0.0606  0.0626  0.0645  0.0639  0.0676
A,KWH2500-4999,NETC,EUR,BA  :    :    :    :    :   0.0381  0.0388  0.0367  0.0370  :
A,KWH2500-4999,NETC,EUR,BE  :    :    :    :    :   0.1055  0.1116  0.1092  0.1049  0.1046
A,KWH2500-4999,NETC,EUR,BG  :    :    :    :    :   0.0232  0.0242  0.0256  0.0265  0.0273
A,KWH2500-4999,NETC,EUR,CY  :    :    :    :    :   0.0313 d    0.0319  0.0320  0.0296  0.0272
A,KWH2500-4999,NETC,EUR,CZ  :    :    :    :    :   0.0483  0.0520  0.0557  0.0534  0.0425
..
```

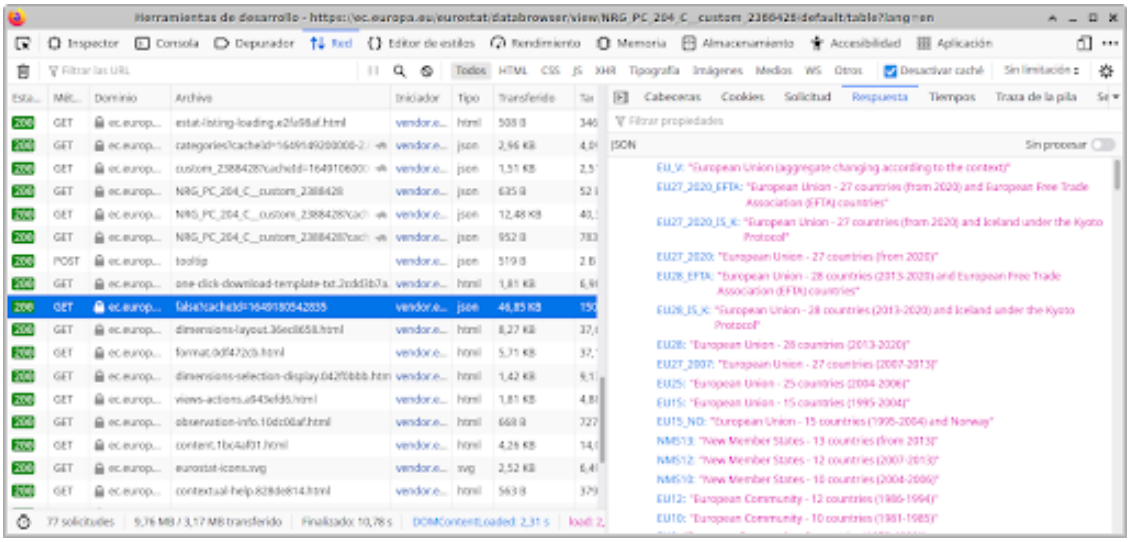
Llamada HTTP para la obtención del los datos de los dataset Analizando la petición anterior, la llamada que se hará en el caso práctico para la obtención de los diferentes dataset que componen el dataset principal, será una petición `GET` HTTP a la URL

`https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/LIVE/false/tsv/<key_dataset>?i`

Donde:

- `<key_dataset>` indica el identificador del dataset del sistema de eurostat, identificado como "*online data code*" en cada uno de los conjuntos de datos que son accesibles desde el "*Data Browser*" de Eurostat. Esa llamada es fácilmente procesable desde las utilidades que ofrece la librería PANDAs para la carga de dataset.

Llamada HTTP para la obtención de los maestros de los países En los conjuntos de datos descargables, los países están establecidos por un identificador. Para obtener el nombre de los países a partir de su identificador, se utiliza la siguiente petición, también obtenida por ingeniería inversa gracias a las herramientas de desarrollo del navegador.



Captura de pantalla de la consola del navegador con la petición del maestro

de datos de los países con respuesta en formato JSON

La llamada HTTP en formato CURL es:

```
$ curl 'https://ec.europa.eu/eurostat/databrowser-backend/api/codelist/LIVE/GEO/
getCodeListJson/9.0/ESTAT/en/false?cacheId=1649180542835'...
```

Como respuesta se obtiene un JSON en cuya estructura en la clave `category/label` donde se encuentra el mapeo entre el identificador y el nombre del país.


```
{
    "version": "2.0",
    "class": "dimension",
    "source": "ESTAT",
    "category": {
        "label": {
            "EUR": "Europe",
            "EU": "European Union (EU6-1958, EU9-1973, EU10-1981, EU12-1986, EU15-1995, EU25-2(
            "EU_V": "European Union (aggregate changing according to the context)",
            "EU27_2020_EFTA": "European Union - 27 countries (from 2020) and European Free Trac
            "EU27_2020_IS_K": "European Union - 27 countries (from 2020) and Iceland under the
            "EU27_2020": "European Union - 27 countries (from 2020)",
            "EU28_EFTA": "European Union - 28 countries (2013-2020) and European Free Trade Ass
            "EU28_IS_K": "European Union - 28 countries (2013-2020) and Iceland under the Kyoto
            "EU28": "European Union - 28 countries (2013-2020)",
            "EU27_2007": "European Union - 27 countries (2007-2013)",
            "EU25": "European Union - 25 countries (2004-2006)",
            "EU15": "European Union - 15 countries (1995-2004)",
...
```


Esta información se obtiene desde el notebook de jupyter gracias a las librerías `requesty` `json`

1.6.2 Dataset auxiliares obtenidos para la generación del dataset principal

Los dataset relacionados con los precios del gas y la electricidad en los países europeos que se han utilizado para crear el dataset principal son:

- **Precios del gas (Euro/kWh) para consumidores domésticos**.(Gas prices components for household consumers - annual data: https://ec.europa.eu/eurostat/databrowser/view/nrg_pc_202_c/default/table?lang=en). Código del dataset (*online data code*): `NRG_PC_202_C`
- **Precios del gas (Euro/kWh) para empresas**.(Gas prices components for non-household consumers - annual data: https://ec.europa.eu/eurostat/databrowser/view/nrg_pc_203_c/default/table?lang=en). Código del dataset (*online data code*): `NRG_PC_203_C`
- **Precio de la electricidad (Euro/kWh) para consumidores domésticos para la banda de consumo entre 2.500 a 4.999 kWh** (Electricity prices components for household consumers - annual data (from 2007 onwards): (https://ec.europa.eu/eurostat/databrowser/view/NRG_PC_204_C__custom_2388428/default/table?lan Código del dataset (*online data code*): `NRG_PC_204_C`

- **Precio de la electricidad (Euro/kWh) para empresas para la banda de consumo de menos de 20 MWh.**(Electricity prices components for non-household consumers - annual data (from 2007 onwards): https://ec.europa.eu/eurostat/databrowser/view/nrg_pc_205_c/default/table?lang=en). Código del dataset (*online data code*): `NRG_PC_205_C`

1.6.3 Procesamiento de los dataset auxiliares

Es necesario realizar un proceso de procedimento de procesamiento y tratamiento de la información en bruto de de los dataset. Todos cumplen la misma estructura por lo que el flujo de tratamiento es el mismo, con los siguientes pasos:

1) Filtrar los datos. Los dataset contienen información adicional que no es necesaria para el alcance de este proyecto. Los filtros comunes a todos los dataset son:
   - Datos anuales
   - Componentes del precio de la energia: *"Energia y suministro"*
   - Moneda: Euro (€)
   - Unidad de medida: kWh
2) Obtener el identificador del país. Este dato viene incluido con más información y es necesario extraerla.
3) Eliminará los espacios de los nombres de las columnas de los dataframe.
4) Procesará todas aquellas columnas relativas a los años para que contengan datos númericos. Algunos datos vienen marcados con indicadores, como dato confidencial y estimado. Se ha tomado la decisión de que los datos confidenciales se tratan como vacíos y los estimados como valores reales.
5) Se añade la columna con la descripción del país.

Filtrado para los precios del gas (Euro/kWh) Para ambos dataset relacionados con el gas, se filtrarán los datos que cumplan los filtros comunes que se han indicado más arriba y además:

- Consumición de la energía: En Giga Júlios en todas las bandas

Filtrado para los precios de la electricidad (Euro/kWh) para consumidores domésticos Se utilizará los filtros comunes y se tendrá en cuenta sólo:

- Consumición de la energía: Consumición entre 2500 kWh y 4999 kWh

Filtrado para los precios de la electricidad (Euro/kWh) para empresas Se utilizará los filtros comunes y se tendrá en cuenta sólo:

- Consumición de la energía: Consumición menos de 20 MWh

1.6.4 Exportación de los dataset auxiliares.

El el repositorio de código de este ejercico existe una exportación en formato CSV de los dataset auxiliares procesados dentro del directorio `subdataset/`

1.6.5 Generación del dataset principal

Después de procesar los datos auxiliares se procede primero a sanitizar y si procede estimar valores. Para ello para dataset auxiliar se hace un estudio de los valores mostrandolos en un gráfico de

caja. En aquellos valores que no están definidos se evaluarán utilizando un estimador de la media. También como proceso de sanitazación se inicializarán aquellos valores *Outlier*.

Por último, para la confencción final del dataset se unén los valores de los diferentes dataframes utilizando como clave el identificador del país, de tal modo que cada registro contenga el indentificador y el nombre del país y los valores de los precios en Eur/kWh del gas y la electricidad tanto de uso doméstico como industrial para los años entre el 2017 y el 2021.

1.7 Campos que incluye el dataset

Nombre del campo	Tipo	Descripción
country	Cadena	El identificador del país
country_name	Cadena	Nombre del país (En inglés)
2017_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2017
2018_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2018
2019_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2019
2020_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2020
2021_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2021
2017_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2017
2018_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2018
2019_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2019
2020_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2020
2021_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2021
2017_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2017
2018_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2018
2019_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2019
2020_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2020
2021_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2021
2017_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2017
2018_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2018

Nombre del campo	Tipo	Descripción
2019_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2019
2020_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2020
2021_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2021

- La periocidad del dato es anual
- Los precios están en Eur/Kwh y corresponden a los costes de la energía y el suministro
- En el caso de la electricidad doméstica es el precio de la banda de consumo entre 2500 kWh y 4999 kWh
- En el caso de la electricidad industrial es el precio de la banda de consumo de menos de 20 MWh

1.7.1 Exportación del dataset principal

El el repositorio de código de este ejercio existe una exportación en formatos CSV del dataset principal en la ruta `dataset/energy_price_dataset.csv`

1.8 Agradecimientos

El propietario de los datos es la Comisión Europea de Estadística y se ha actuado en base al cometido de uso, distribución y potenciación de datos estadísticos oficiales de cada país perteneciente a la zona Euro, su finalidad es divulgativa. Otros análisis que lo referencia:

- https://pdfs.semanticscholar.org/8472/6fef650b500f5928993bd6a0b2395429474e.pdf
- https://cincodias.elpais.com/cincodias/2022/04/06/companias/1649271023_196237.html

El tipo de licenciamiento de los datos es:

Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

1.9 Inspiración

El juego de datos que hemos decidido extraer, tiene una especial relevancia actual, debido a la importancia de los resultados que podemos obtener de su uso en análisis, debido que en el panorama actual, el poder realizar estudios que arrojen información acerca de la evolución económica de los mercados energéticos, cobra una especial relevancia.

Se pueden obtener conclusiones, después de todo el proceso de extracción, depuración y enriquecimiento, bastante interesantes. Como por ejemplo, que muchos de los países con mejor índice económico de la zona Euro, tienen precios más competitivos que otros países con peor indice económico. En análisis anteriores, usan datasets diferentes para análisis similares y objetivos parecidos, tal y como se muestra en el apartado 6.

1.10 Licencia

Obtamos por el mismo tipo de licenciamiento que el proveedor de los datos extraidos, ya que como cuya finalidad es divulgativa, permitimos su uso, incluso comercial y su manipulación/transformación:

Released Under CC BY-SA 4.0 License.

1.11 Código

El código de está práctica está incluido en el siguiente repositorio de Github:

https://github.com/tipologia-datos-UOC-pract-jonas-miguel/pract-1

Para el desarrollo de este ejercicio práctico se ha utilizado un notebook de Python de jupyter) que puede ser ejecutado a través de docker. El archivo README.md del repositorio de código contine información de como poder arrancar este servicio.

El notebook se encuentra en la siguiente ruta:

- `notebook/pract-1.ipynb`

1.12 Dataset

El dataset se encuentra publicado en Zenodo, DOI: `10.5281/zenodo.6424152` **(https://doi.org/10.5281/zenodo.6424152)**

1.13 Firmas

Contribuciones	Firma
Investigación previa	**Jonás Medina Brito** (jmedinabrit@uoc.edu), **Miguel Rafael Esteban Martín** (mestebanmart@uoc.edu)
Redacción de las respuestas	**Miguel Rafael Esteban Martín** (mestebanmart@uoc.edu), **Jonás Medina Brito** (jmedinabrit@uoc.edu)
Desarrollo del código	**Miguel Rafael Esteban Martín** (mestebanmart@uoc.edu), **Jonás Medina Brito** (jmedinabrit@uoc.edu)