

doc-pract-1

April 7, 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 Índice de contenidos

- Contexto
- Título
- Descripción del dataset
- Representación gráfica
- Contenido
- Agradecimientos
- Inspiración
- Licencia
- Código
- Dataset

1.2 Contexto

1.2.1 Recogida de datos

Aunque [Eurostat](#) tiene una API pública (1) que permite obtener los diferentes dataset, en esta práctica hemos decidido 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 \(from 2007 onwards\)](#))

The screenshot shows the Eurostat Data Browser interface. The title is "Electricity prices components for household consumers - annual data (from 2007 onwards)". The data code is "NRG_PC_204_C". The interface includes a selection panel with "Geographical entity (reporting)" set to "EU-28 (EU-27 + EU-28)" and "Time" set to "Year [Yr]". The table displays electricity price components for various European countries from 2007 to 2011.

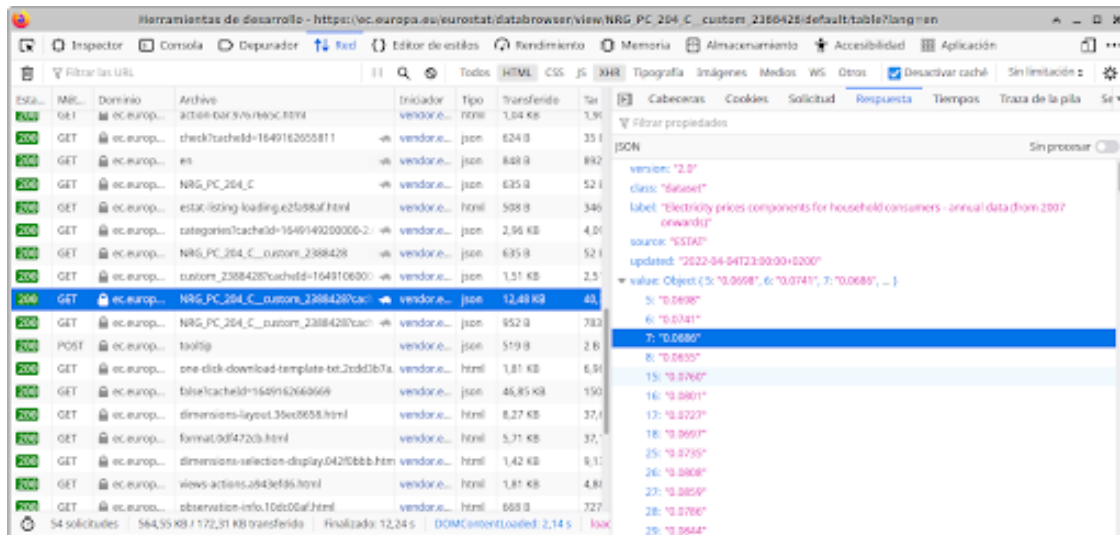
Geographical entity (reporting)	2007	2008	2009	2010	2011
European Union - 27 countries (from 2007)	0.0708	0.0695	0.0700	0.0706	0.0711
EU-28 (EU-27 + EU-28)	0.0708	0.0697	0.0703	0.0703	0.0704
Belgium	0.0700	0.0700	0.0694	0.0693	0.0704
Bulgaria	0.0671	0.0661	0.0664	0.0671	0.0684
Croatia	0.0602	0.0599	0.0593	0.0597	0.0604
Cyprus	0.0678	0.0680	0.0683	0.0682	0.0687
Germany (incl. HH service territory of the FRG)	0.0668	0.0668	0.0670	0.0671	0.0670
Estonia	0.0602	0.0600	0.0600	0.0604	0.0603
Finland	0.0700	0.0701	0.0700	0.0707	0.0708
France	0.0700	0.0696	0.0698	0.0699	0.0699
Greece	0.0600	0.0600	0.0600	0.0600	0.0600
Spain	0.0600	0.0600	0.0600	0.0600	0.0600
France	0.0700	0.0696	0.0698	0.0699	0.0699
Croatia	0.0600	0.0600	0.0600	0.0600	0.0600
Italy	0.0700	0.0696	0.0698	0.0699	0.0699
Poland	0.0600	0.0600	0.0600	0.0600	0.0600
Lithuania	0.0600	0.0600	0.0600	0.0600	0.0600
Latvia	0.0600	0.0600	0.0600	0.0600	0.0600
Luxembourg	0.0600	0.0600	0.0600	0.0600	0.0600
Hungary	0.0600	0.0600	0.0600	0.0600	0.0600
Malta	0.0600	0.0600	0.0600	0.0600	0.0600
Netherlands	0.0600	0.0600	0.0600	0.0600	0.0600
Austria	0.0600	0.0600	0.0600	0.0600	0.0600
Portugal	0.0600	0.0600	0.0600	0.0600	0.0600
Romania	0.0600	0.0600	0.0600	0.0600	0.0600
Slovenia	0.0600	0.0600	0.0600	0.0600	0.0600
Slovakia	0.0600	0.0600	0.0600	0.0600	0.0600
Sweden	0.0600	0.0600	0.0600	0.0600	0.0600
Switzerland	0.0600	0.0600	0.0600	0.0600	0.0600
Turkey	0.0600	0.0600	0.0600	0.0600	0.0600
United Kingdom	0.0600	0.0600	0.0600	0.0600	0.0600
Montenegro	0.0600	0.0600	0.0600	0.0600	0.0600

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.

Utilizando ingeniería inversa y las herramientas de desarrollo del navegador (Firefox), se puede ver

cual es la llamada que devuelve los datos que se representan en la tabla dinámica. Un ejemplo es la petición que es utilizada en la construcción de la tabla con los datos que se muestra en la página web.



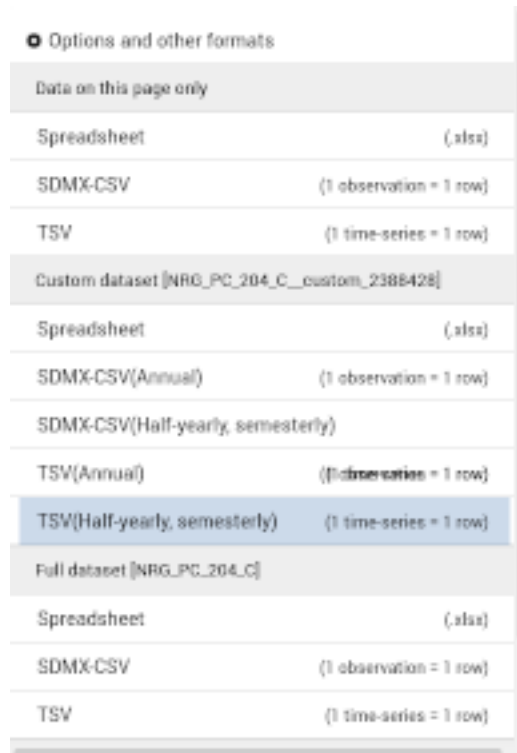
Captura de pantalla de la consola del navegador con la petición de datos de dataset y respuesta JSON

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' ...
```

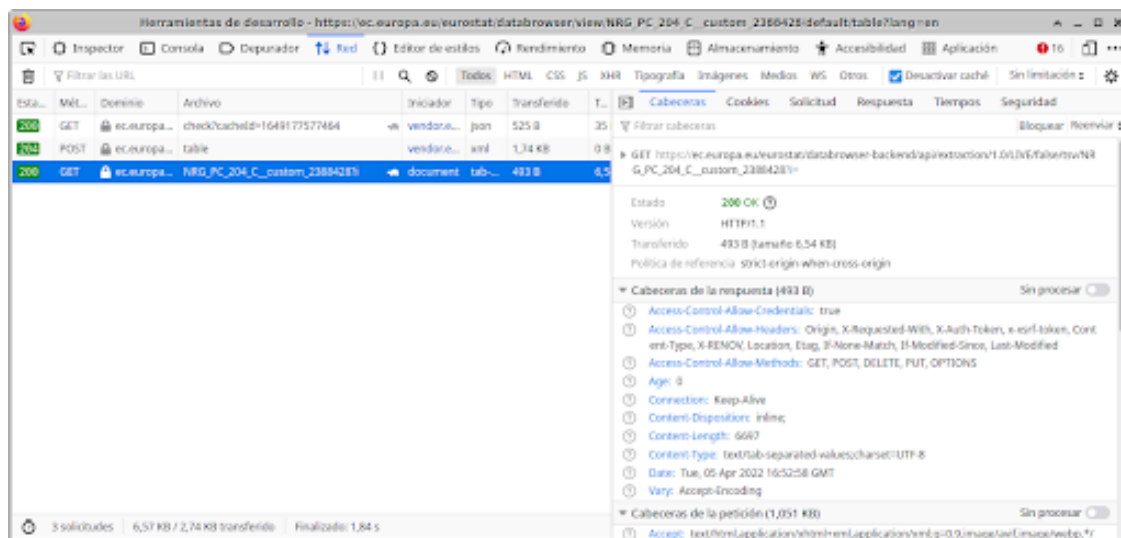
La respuesta de esta llamada es en formato JSON con una estructura específica, difícil de procesar y aprovechar para nuestro propósito de acceder a los datos.

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/
```

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

```

%ipsec%10%dir -s -c -u https://ec.europa.eu/eurostat/databrowser/backends/api/extract?unit=0&live=false&csv=0&_PC_004_C_0_custom_330042071' | gzip
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 6697 100 6697 0 0 43594 0 0 0 0 0 43771
freq,reg,css,arg,prc,currency,geo,TIME_PERIOD
2012-52 2013-52 2014-52 2015-52 2016-52 2017 2018 2019 2020 2021
x_KMG568-4999.NETC,ELR, : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,AT : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,BA : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,BE : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,BG : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,CY : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,CZ : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,DE : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,DK : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,EE : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,EL : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,ES : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,FI : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,FR : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,GB : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,GR : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,HU : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,IE : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,IT : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,LB : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,LI : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,LT : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,LU : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,LV : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,MT : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,NL : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,NO : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,PL : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,PT : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,RO : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,SE : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,SI : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,SK : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,TR : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
x_KMG568-4999.NETC,ELR,UK : : : 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

```

Captura de pantalla de la respuesta utilizando el comando *CURL*

1.2.2 Llamada HTTP para la obtención de los datos de los dataset

Analizando la petición anterior la llamada que se hará para la obtención de los diferentes dataset que componen el dataset principal será una petición GET 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.

- 1.3 Título
- 1.4 Descripción del dataset
- 1.5 Representación gráfica
- 1.6 Contenido
- 1.7 Agradecimientos
- 1.8 Inspiración
- 1.9 Licencia
- 1.10 Código
- 1.11 Dataset