

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)", "Time" set to "Year", and "Page" set to "Time frequency: [Y]". The table below shows the data for various countries, with columns for the year and the price component.

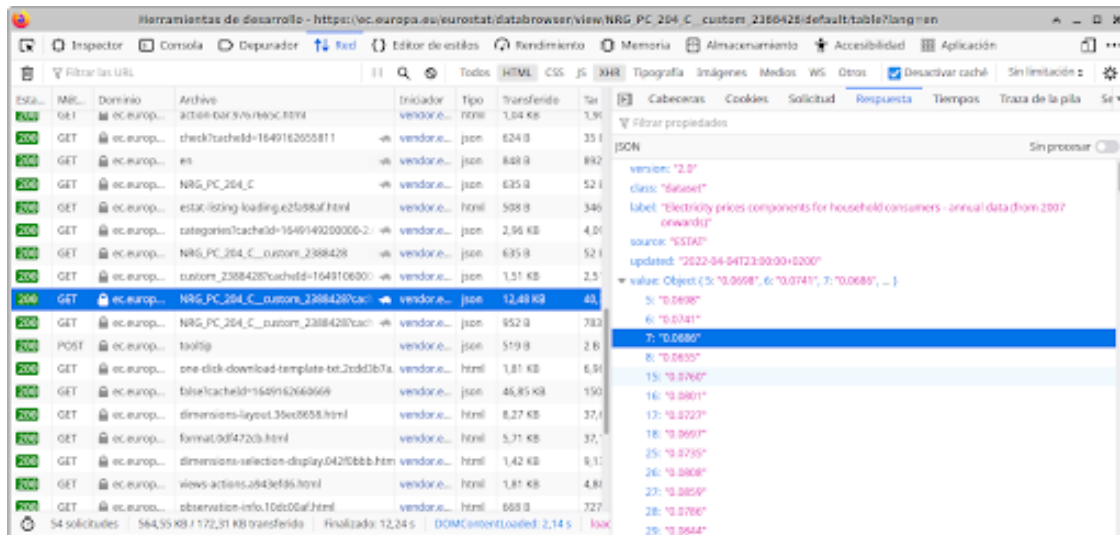
Geographical entity (reporting)	2007	2008	2009	2010	2011
European Union - 27 countries (from 2007)	0.0706	0.0695	0.0700	0.0706	0.0711
European Union - 28 countries (from 2007)	0.0706	0.0697	0.0703	0.0703	0.0704
Austria	0.0700	0.0700	0.0694	0.0693	0.0694
Belgium	0.0671	0.0661	0.0664	0.0671	0.0664
Czechia	0.0602	0.0599	0.0593	0.0597	0.0594
Denmark	0.0578	0.0580	0.0583	0.0582	0.0581
Germany (incl. HH service territory of the FRG)	0.0604	0.0604	0.0610	0.0611	0.0610
Greece	0.0402	0.0407	0.0404	0.0404	0.0404
Ireland	0.0700	0.0701	0.0700	0.0701	0.0700
France	0.0700	0.0696	0.0694	0.0694	0.0694
Spain	0.0602	0.0602	0.0604	0.0604	0.0604
Finland	0.0700	0.0694	0.0697	0.0693	0.0693
Croatia	0.0604	0.0601	0.0591	0.0591	0.0594
Slovenia	0.0700	0.0694	0.0691	0.0693	0.0693
Poland	0.0604	0.0604	0.0604	0.0604	0.0604
Lithuania	0.0604	0.0604	0.0604	0.0604	0.0604
Latvia	0.0604	0.0604	0.0604	0.0604	0.0604
Malta	0.0604	0.0604	0.0604	0.0604	0.0604
Portugal	0.0604	0.0604	0.0604	0.0604	0.0604
Romania	0.0604	0.0604	0.0604	0.0604	0.0604
Slovakia	0.0604	0.0604	0.0604	0.0604	0.0604
Hungary	0.0604	0.0604	0.0604	0.0604	0.0604
Netherlands	0.0604	0.0604	0.0604	0.0604	0.0604
Sweden	0.0604	0.0604	0.0604	0.0604	0.0604
Italy	0.0604	0.0604	0.0604	0.0604	0.0604
United Kingdom	0.0604	0.0604	0.0604	0.0604	0.0604
Midway	0.0604	0.0604	0.0604	0.0604	0.0604

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.



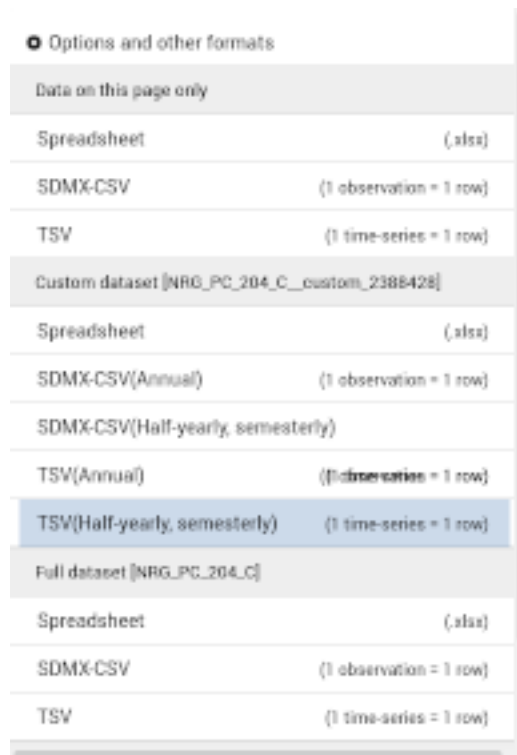
Captura de pantalla de la consola del navegador con la petición de datos de dataset y respuesta JSON

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' ...
```

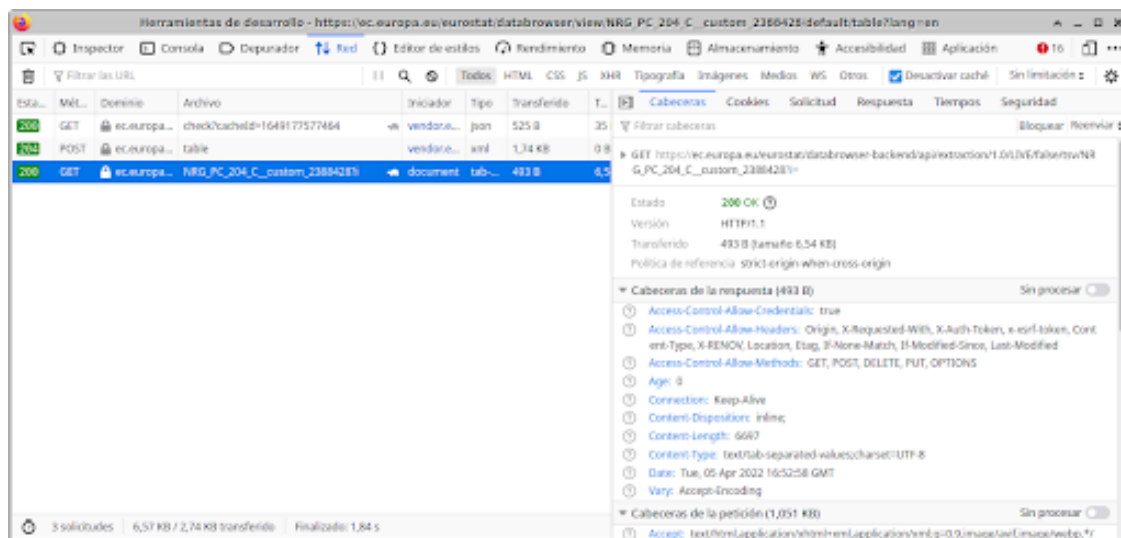
La respuesta de esta llamada es en formato JSON con una estructura específica, difícil de procesar y aprovechar para nuestro propósito de acceder a los datos.

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:



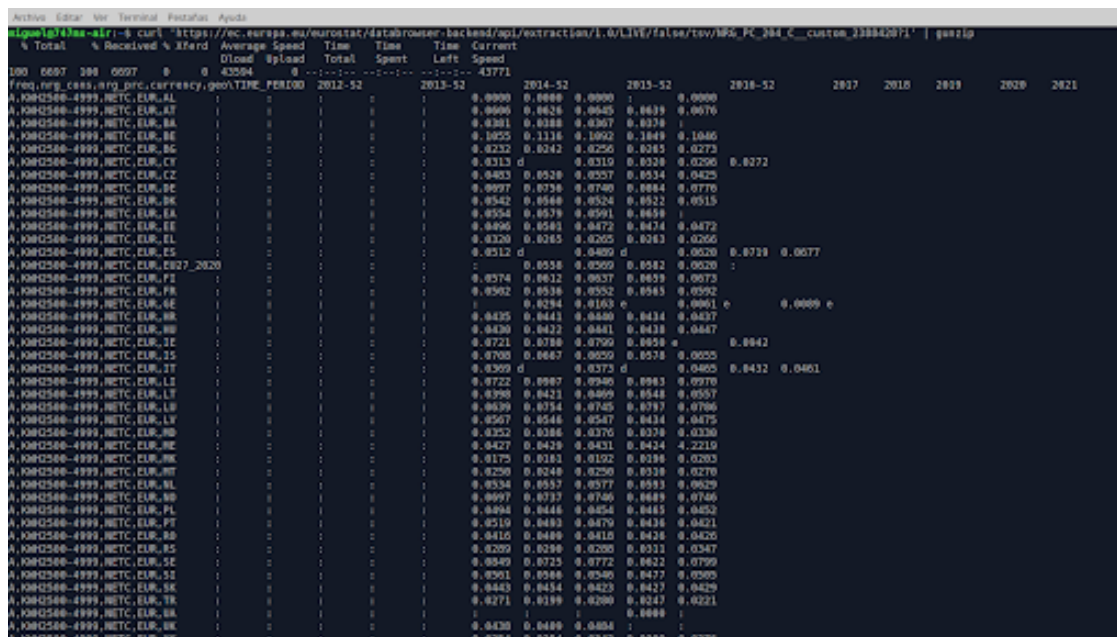
Captura de pantalla de la interfaz “Data Browser” de la descarga del informe de los componentes del precio de la electricidad para consumidores domésticos en formato TSV

Qué si se exporta la petición en formato CURL es:

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

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



```
Active Editor: Vim Terminal: PuTTY: Apache
niguel@4hms-airi:~$ curl -s https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/LIVE/false/tsv/NRG_PC_204_C__custom_2388428?i | gunzip
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           % Done    0         0             0      0 --:--:-- --:--:-- --:--:-- 0
100 669T 300 669T    0 43594 0 --:--:-- --:--:-- --:--:-- 43771
freq,nrg_cas,nrg_prc,carrecncy,geo,TIME_PERIOD 2012-52 2013-52 2014-52 2015-52 2016-52 2017 2018 2019 2020 2021
A,KHQ2580-4999,NETC,EUR,AL 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
A,KHQ2580-4999,NETC,EUR,AT 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
A,KHQ2580-4999,NETC,EUR,BA 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
A,KHQ2580-4999,NETC,EUR,BE 0.1055 0.1116 0.1092 0.1049 0.1046 0.1046 0.1046 0.1046 0.1046 0.1046
A,KHQ2580-4999,NETC,EUR,BG 0.0232 0.0242 0.0256 0.0265 0.0273 0.0273 0.0273 0.0273 0.0273 0.0273
A,KHQ2580-4999,NETC,EUR,CY 0.0313 d 0.0319 0.0320 0.0296 0.0272 0.0272 0.0272 0.0272 0.0272 0.0272
A,KHQ2580-4999,NETC,EUR,CC 0.0403 0.0520 0.0537 0.0534 0.0425 0.0425 0.0425 0.0425 0.0425 0.0425
A,KHQ2580-4999,NETC,EUR,DE 0.0027 0.0736 0.0740 0.0804 0.0770 0.0770 0.0770 0.0770 0.0770 0.0770
A,KHQ2580-4999,NETC,EUR,DK 0.0542 0.0540 0.0524 0.0522 0.0515 0.0515 0.0515 0.0515 0.0515 0.0515
A,KHQ2580-4999,NETC,EUR,EA 0.0554 0.0579 0.0591 0.0609 0.0609 0.0609 0.0609 0.0609 0.0609 0.0609
A,KHQ2580-4999,NETC,EUR,EE 0.0496 0.0501 0.0472 0.0474 0.0472 0.0472 0.0472 0.0472 0.0472 0.0472
A,KHQ2580-4999,NETC,EUR,EL 0.0320 0.0265 0.0205 0.0263 0.0266 0.0266 0.0266 0.0266 0.0266 0.0266
A,KHQ2580-4999,NETC,EUR,ES 0.0512 d 0.0489 d 0.0520 0.0520 0.0520 0.0520 0.0520 0.0520 0.0520 0.0520
A,KHQ2580-4999,NETC,EUR,EU27_2020 0.0558 0.0509 0.0582 0.0620 0.0620 0.0620 0.0620 0.0620 0.0620 0.0620
A,KHQ2580-4999,NETC,EUR,FI 0.0374 0.0612 0.0637 0.0659 0.0673 0.0673 0.0673 0.0673 0.0673 0.0673
A,KHQ2580-4999,NETC,EUR,FR 0.0502 0.0536 0.0552 0.0565 0.0592 0.0592 0.0592 0.0592 0.0592 0.0592
A,KHQ2580-4999,NETC,EUR,GE 0.0294 0.0163 e 0.0001 e 0.0000 e 0.0000 e 0.0000 e 0.0000 e 0.0000 e 0.0000 e
A,KHQ2580-4999,NETC,EUR,GR 0.0435 0.0441 0.0450 0.0434 0.0437 0.0437 0.0437 0.0437 0.0437 0.0437
A,KHQ2580-4999,NETC,EUR,HU 0.0430 0.0422 0.0441 0.0438 0.0447 0.0447 0.0447 0.0447 0.0447 0.0447
A,KHQ2580-4999,NETC,EUR,IE 0.0721 0.0780 0.0790 0.0808 0.0808 0.0808 0.0808 0.0808 0.0808 0.0808
A,KHQ2580-4999,NETC,EUR,IS 0.0700 0.0667 0.0650 0.0578 0.0655 0.0655 0.0655 0.0655 0.0655 0.0655
A,KHQ2580-4999,NETC,EUR,IT 0.0399 d 0.0373 d 0.0405 0.0432 0.0461 0.0461 0.0461 0.0461 0.0461 0.0461
A,KHQ2580-4999,NETC,EUR,LI 0.0722 0.0687 0.0640 0.0663 0.0670 0.0670 0.0670 0.0670 0.0670 0.0670
A,KHQ2580-4999,NETC,EUR,LT 0.0398 0.0421 0.0409 0.0548 0.0557 0.0557 0.0557 0.0557 0.0557 0.0557
A,KHQ2580-4999,NETC,EUR,LU 0.0639 0.0734 0.0745 0.0797 0.0796 0.0796 0.0796 0.0796 0.0796 0.0796
A,KHQ2580-4999,NETC,EUR,LV 0.0567 0.0546 0.0547 0.0434 0.0475 0.0475 0.0475 0.0475 0.0475 0.0475
A,KHQ2580-4999,NETC,EUR,MD 0.0352 0.0386 0.0370 0.0378 0.0330 0.0330 0.0330 0.0330 0.0330 0.0330
A,KHQ2580-4999,NETC,EUR,ME 0.0427 0.0429 0.0421 0.0424 0.2210 0.2210 0.2210 0.2210 0.2210 0.2210
A,KHQ2580-4999,NETC,EUR,NL 0.0175 0.0161 0.0192 0.0196 0.0203 0.0203 0.0203 0.0203 0.0203 0.0203
A,KHQ2580-4999,NETC,EUR,NO 0.0250 0.0240 0.0250 0.0310 0.0270 0.0270 0.0270 0.0270 0.0270 0.0270
A,KHQ2580-4999,NETC,EUR,PL 0.0334 0.0557 0.0577 0.0593 0.0629 0.0629 0.0629 0.0629 0.0629 0.0629
A,KHQ2580-4999,NETC,EUR,PT 0.0697 0.0737 0.0746 0.0688 0.0746 0.0746 0.0746 0.0746 0.0746 0.0746
A,KHQ2580-4999,NETC,EUR,RO 0.0494 0.0446 0.0454 0.0415 0.0452 0.0452 0.0452 0.0452 0.0452 0.0452
A,KHQ2580-4999,NETC,EUR,RS 0.0510 0.0483 0.0479 0.0436 0.0421 0.0421 0.0421 0.0421 0.0421 0.0421
A,KHQ2580-4999,NETC,EUR,SE 0.0410 0.0440 0.0415 0.0426 0.0426 0.0426 0.0426 0.0426 0.0426 0.0426
A,KHQ2580-4999,NETC,EUR,SI 0.0209 0.0290 0.0200 0.0311 0.0347 0.0347 0.0347 0.0347 0.0347 0.0347
A,KHQ2580-4999,NETC,EUR,SK 0.0840 0.0725 0.0772 0.0622 0.0799 0.0799 0.0799 0.0799 0.0799 0.0799
A,KHQ2580-4999,NETC,EUR,SV 0.0501 0.0500 0.0540 0.0477 0.0509 0.0509 0.0509 0.0509 0.0509 0.0509
A,KHQ2580-4999,NETC,EUR,TR 0.0443 0.0454 0.0422 0.0427 0.0429 0.0429 0.0429 0.0429 0.0429 0.0429
A,KHQ2580-4999,NETC,EUR,UA 0.0271 0.0199 0.0200 0.0247 0.0221 0.0221 0.0221 0.0221 0.0221 0.0221
A,KHQ2580-4999,NETC,EUR,UK 0.0430 0.0440 0.0404 0.0404 0.0404 0.0404 0.0404 0.0404 0.0404 0.0404
A,KHQ2580-4999,NETC,EUR,XX 0.0364 0.0354 0.0354 0.0359 0.0388 0.0388 0.0388 0.0388 0.0388 0.0388
```

Captura de pantalla de la respuesta utilizando el comando CURL

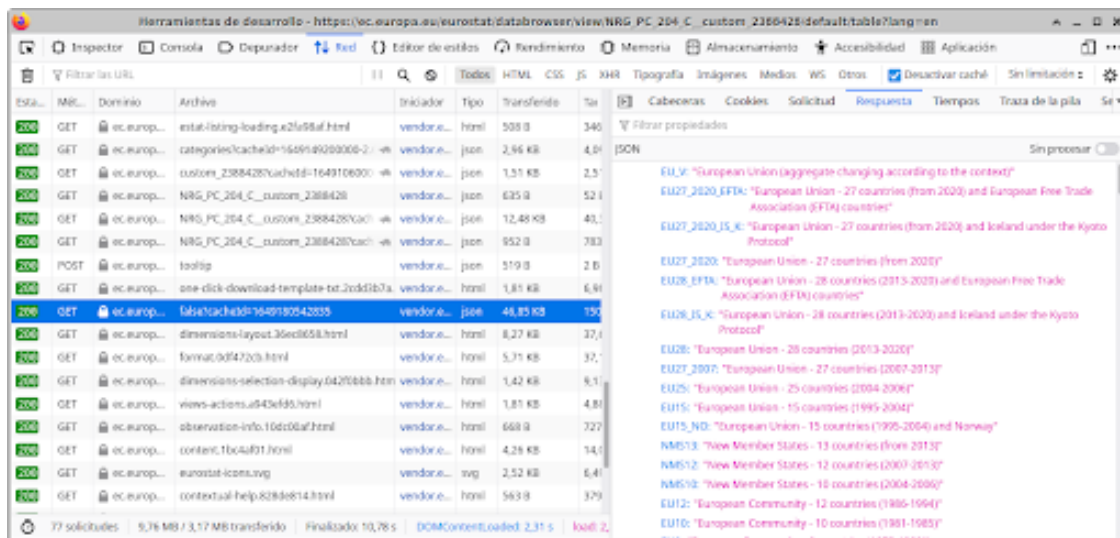
Llamada HTTP para la obtención de los datos de los dataset Analizando la petición anterior, la llamada que se hará en el caso práctico para la obtención de los diferentes dataset que componen el dataset principal, será una petición GET HTTP a la URL

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

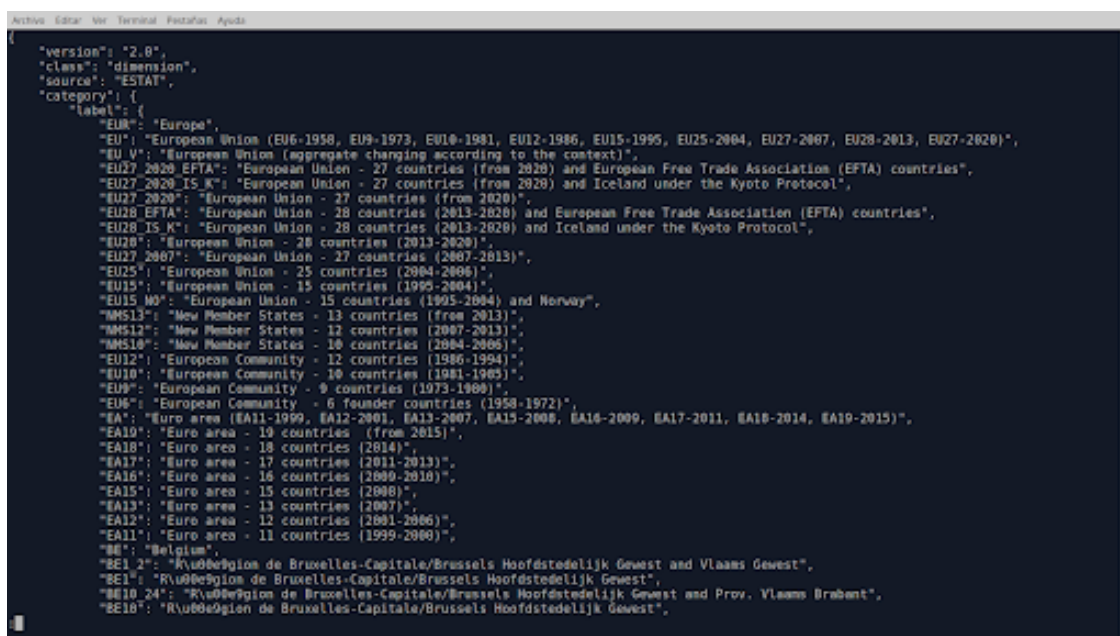


Captura de pantalla de la consola del navegador con la petición del maestro de datos de los países con respuesta en formato JSON

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.



Captura de pantalla de la respuesta en formato JSON de la petición del maestro de categorías.

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

1.2.2 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 para consumidores domésticos. ([Gas prices components for household consumers - annual data](#)). Código del dataset (*online data code*): **NRG_PC_202_C**
- Precios del gas 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 para consumidores domésticos para la banda de consumo entre 2.500 a 4.999 kWh ([Electricity prices components for household consumers - annual data \(from 2007 onwards\)](#)). Código del dataset (*online data code*): **NRG_PC_204_C**
- Precio de la electricidad para empresas. ([Electricity prices components for non-household consumers - annual data \(from 2007 onwards\)](#)). Código del dataset (*online data code*): **NRG_PC_205_C**

Procesamiento de los dataset auxiliares Es necesario realizar un proceso de procedimiento 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:

- 1) Filtrar los datos. El dataset contiene información adicional que no es necesaria para el alcance de este proyecto.
- 2) Obtener el identificador del país Este dato viene incluido con más información y es necesario extraerla
- 3) Eliminar los espacios de los nombres de las columnas de los dataframe.
- 4) Procesará todas aquellas columnas relativas a los años para que contengan datos numéricos. Algunos datos vienen marcados con flags, 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.

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