

1. Context. Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació. Indicar l'adreça del lloc web.

El projecte que a continuació es presenta té per objectiu recollir dades sobre els preus de l'energia en moneda euro del megavat/hora (en endavant, MWh) durant un període seleccionat de dates (1 de Novembre del 2020 al 31 d'Octubre del 2022). A més, també recull dades de la producció diària provinent de fonts d'energia renovables en megavats (en endavant, MW) en el mateix rang de dates.

El tema del projecte va ser escollit pel integrants del grup pel seu interès personal i la seva rellevància en l'economia actual. Tal i com es menciona en diferents mitjans de comunicació, la situació originada per la guerra a Ucraïna i l'especulació de les empreses energètiques està generant un augment en el preu de l'energia mai vist fins ara. [1, 2] De retruc, això està generant un augment de preus en altres àmbits, provocant una situació d'inflació insostenible per a moltes famílies i comerços. [1, 3]

D'altra banda, hi ha una necessitat urgent d'assolir una completa transició energètica a fonts d'energia renovables que eliminin la dependència dels combustibles fòssils i l'augment perjudicial en les emissions de gasos d'efecte hivernacle que estan causant el canvi climàtic antropogènic. Així doncs, l'augment de la producció d'energia de fonts renovables hauria de contribuir a una reducció dels preus de l'energia.

Malgrat que diversos llocs web es van considerar en un inici per tal de portar a terme aquest projecte, la decisió final d'utilitzar "Red Eléctrica - ESIOS (Sistema de Información del Operador del Sistema) es va prendre degut a què és l'operador que gestiona la xarxa elèctrica a l'Estat espanyol i el generador de les dades. Per tant, el seu lloc web conté les dades necessàries per a dur a terme el projecte. D'una banda, presenta les dades del preu de l'energia amb una granularitat horària. De l'altra, aporta informació amb la mateixa granularitat de la producció d'origen renovable i per tipologia (eòlica, solar, hidroelèctrica, etc.). A més a més, també presenta les característiques que incrementen la dificultat de realització de la pràctica, com ara la possibilitat de portar a terme l'*scraping* amb tecnologies avançades com Selenium, l'ús de contingut dinàmic dels gràfics o la gestió de codi Javascript al moment de l'extracció de dades.

Finalment, l'adreça del lloc web s'adjunta a continuació: <https://www.esios.ree.es/es/>. En particular s'han extret dades de les subpàgines: **Mercados y precios** i **Generación y consumo**. Cal mencionar que aquesta pàgina disposa d'una API per accedir a les dades, però tal i com s'explicita a la pràctica, però no se n'ha fet ús ja que les dades faltants eren mínimes.

2. Títol. Definir un títol que sigui descriptiu pel dataset.

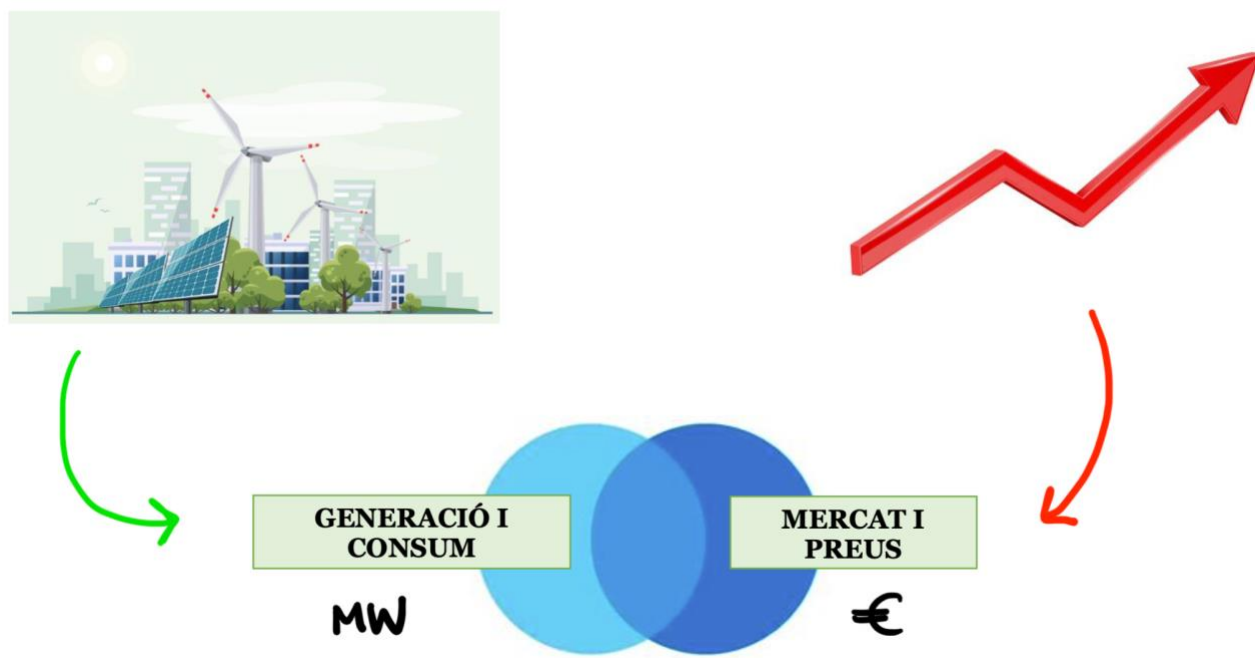
Estudi de l'evolució del preu de l'electricitat en funció de la producció d'energia renovable.

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.

El dataset obtingut recull diferents tipus de dades per estudiar l'evolució dels preus de l'energia durant un període d'aproximadament un any entre el 1/11/2020 i el 31/10/2022. En concret, s'han recollit dades de les dates i hores dels preus promitjos de l'energia (en MWh) totals i en funció del tipus de mercat (lliure i de referència), del volum d'energia produït. També s'han obtingut dades de la proporció d'energia renovable per tipus de font (eòlica, solar, hidroelèctrica, etc.) durant el mateix període i amb la mateixa granularitat horària.

El format del dataset es guardat en un fitxer CSV que facilita el seu posterior anàlisi, visualització i tractament de neteja i preprocessat de les dades.

4. Representació gràfica. Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.



5. Contingut. Explicar els camps que inclou el dataset i el període de temps de les dades.

Com s'ha mencionat anteriorment, el dataset inclou dades en un període del 1/11/2020 al 31/10/2022. Més concretament, la definició dels camps que s'extreuen i formen part del dataset són:

- **Date:** data de l'obtenció del registre amb el format aaaa-mm-dd.
- **Hora:** hora de l'obtenció del registre amb el format hh:mm (24 h).
- **Avg total price (euro/MWh):** preu total promig del megavat hora.
- **Avg price free market (euro/MWh):** preu promig del megavat hora pel comercialitzador lliure.
- **Avg price reference market (euro/MWh):** preu promig del megavat hora pel comercialitzador de referència.
- **Energy total (MWh):** volum total d'energia produïda.
- **Energy free market (MWh):** volum total d'energia produïda en el mercat lliure.
- **Energy reference market (MWh):** volum total d'energia produïda en el mercat de referència.
- **Free market share (%):** proporció del mercat lliure en el volum d'energia produïda.
- **Reference market share (%):** proporció del mercat de referència en el volum d'energia produïda.
- **Renewable generation (%):** percentatge de producció d'energia renovable.

Adicionalment, mitjançant una transformació a variable categòrica, s'afegeixen les següents dues columnes al dataset final:

- **Energy source:** contribució de la font d'energia a la producció renovable (total renovable, eòlica, solar, hidroelèctrica, nuclear i tèrmica renovable).
- **Energy generation (MW):** producció d'energia renovable en megavats.

Més concretament, les dades del preu de l'energia (Avg total price, Avg price free market i Avg price reference market), volum d'energia (Energy total, Energy free market i Energy reference market) i la quota que representen (Free market share i Reference market share) es recullen de "<https://www.esios.ree.es/es/mercados-y-precios>" (**Fig. 2**) i les dades del percentatge de generació d'energia renovable (Renewable generation), les fonts d'energia (Energy source) i la potencia generada (Energy generation) (**Fig. 3**) es recullen de "<https://www.esios.ree.es/es/generacion-y-consumo>". Un exemple de les taules d'on s'han extret les dades es mostra a continuació:

	TOTAL	COMERCIALIZADOR LIBRE	COMERCIALIZADOR DE REFERENCIA
PRECIO MEDIO €/MWh	110,61	110,61	110,67
ENERGÍA MWh	27.583,2	25.567,7	2.015,5
CUOTA %		92,7	7,3

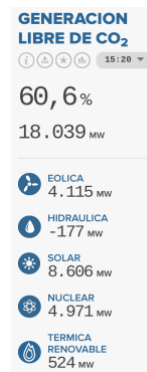


Fig 2. Exemple de dades extretes de la pàgina **Mercados y precios.**

Fig 3. Exemple de dades extretes de la pàgina **Generación y consumo.**

6. Propietari. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

Descripció del propietari de les dades:

El propietari del lloc web i del conjunt de dades és l'empresa Red Eléctrica de España. Aquesta corporació es qui facilita les dades en format web i gràfic. L'empresa es propietaria de les dades i gestiona els serveis generals d'energia a l'estat espanyol.

Anàlisis similars:

S'ha realitzat una cerca d'anàlisis similars on s'han trobat estudis de l'evolució del preu de l'electricitat en funció del temps. Aquestes estudien l'influència que ha tingut el topall al preu del gas en el preu del MWh i l'evolució del preu respecte altres països de la Unió Europea, com també l'evolució de les emissions de CO₂. [4-6] Altres fonts consultades analitzen l'evolució temporal de la demanda elèctrica i de la producció d'energia renovable [7]. Tot i això, cap de les fonts consultades ha analitzat si hi ha una relació entre el preu del MWh i la producció d'energia renovable dins el mateix rang i geolocalització de les dades que el nostre.

Actuacions d'acord amb principis ètics i legals:

Per poder dur a terme el projecte amb el lloc web esmentat anteriorment, es va consultar el fitxer robots.txt (<https://www.ree.es/robots.txt>) per tal d'aclarir si es permetia l'ús d'eines de webscraping per a obtenir les dades necessàries. Tanmateix, per estar segurs que no s'incorria en cap problema legal, l'empresa ha estat contactada directament a través d'un formulari web

(<https://www.ree.es/es/digame>) on s'ha demanat permís explícit per l'ús de *web scraping* per a la realització de la pràctica.

A continuació, s'adjunta la resposta al correu electrònic:

----- Forwarded message -----

De: Digame <digame@ree.es>
Date: dv., 4 de nov. 2022 a les 12:18
Subject: RV: Consulta o petición - Webmaster Ref..DIPT2265838
To: emanzanoma@uoc.edu <emanzanoma@uoc.edu>

Estimada/o Sra./Sr. Manzano y Queraltó,
Nuestros técnicos nos indican que dada la finalidad, pueden utilizar los datos que necesitan bajo las siguientes condiciones:

- Mencionar de forma clara y explícita a Red Eléctrica de España como fuente de la información y de los datos.
- No alterar ni desnaturalizar el sentido de la información ni de los datos.
- No se podrá sugerir o indicar que el titular de la información patrocina o apoya la reutilización que se realice.
- Se prohíbe expresamente la utilización de estos datos con fines comerciales.

Reciba un cordial saludo,

red eléctrica

Servicio DÍGAME
Atención a grupos de interés externos
Dpto. de Sostenibilidad
Dirección de Sostenibilidad
Pº del Conde de los Gaitanes, 177
28109 Alcobendas (Madrid)
Tlf:+34 91 728 62 15
Fax: 91 650 45 42 / 76 77
digame@ree.es
www.ree.es



Fig 4. Comunicació per correu electrònic amb Red Eléctrica de España on se'ns permet extreure dades mitjançant l'ús de web scraping per a la realització de la pràctica.

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

El conjunt de dades generat ens aporta informació sobre de l'evolució dels preus de l'energia i la generació d'energia renovable permetent-nos avaluar la seva evolució temporal i, per tant, preveure quina podria ser l'evolució futura, fet que la fa interessant i alhora útil pel consumidor de les dades. La dependència temporal dels preus de l'electricitat i de la generació renovable ha estat estudiada amb anterioritat, tal i com s'ha explicat en l'apartat anterior. [4-7] Això és especialment important degut a l'impacte global del preu de l'electricitat i les energies, així com

també sobre l'impacte econòmic en diferents àmbits originat per les fluctuacions en la generació i, sobretot, dels seus preus.

Adicionalment, la font de les energies són un pilar essencial quan es parla de sostenibilitat i medi ambient, fet pel qual ens va semblar rellevant incloure aquesta variable en el dataset.

Així doncs, a més d'estudiar l'evolució temporal com ja s'ha fet en altres estudis, el nostre objectiu és buscar una relació més directa entre la fluctuació en la generació d'energia renovable i el preu del MWh. Finalment, la idea seria obtenir un model predictiu que permeti conèixer el preu del MWh en funció dels paràmetres de generació d'energia renovable.

Concretament, les preguntes que es volen respondre són les següents:

- Com fluctua el preu de l'energia diàriament? Es manté la relació de preus/hora en diferents dies? És a dir, el preu a una hora específica respecte el màxim diari és el mateix?
- Influeix el tipus de mercat a l'hora de definir el preu de l'energia?
- Quines són les variables més importants que influeixen en el preu de l'energia? (relació mercat lliure i de referència, volum produït, etc)
- Quin és el preu de l'energia per font i, tenint el compte el component medi ambiental i sostenible, quines energies ens serien més adequades?
- Quina influència té la generació d'energia renovable en el preu de l'energia? A nivell global i pel tipus de font renovable?

8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció. Exemples de llicències que poden considerar-se:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Altres (especificar quina).

La llicència seleccionada pel dataset resultant és la següent: **Released Under CC BY-NC-SA 4.0 License.** L'elecció d'aquesta llicència es basa en els requeriments establerts per Red Eléctrica de España on es demana explícitament que les dades no siguin desnaturalitzades. Ahora també es prohibeix l'ús de les dades amb finalitat comercial.

9. Codi. Codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

- Al document PDF s'han de comentar els aspectes més rellevants sobre com el codi realitza el procés de recollida de dades, quines dificultats presenta el lloc web triat, i com les heu resolt
- El codi haurà de situar-se a la carpeta **/source** del repositori.
- S'han d'indicar les llibreries i versions utilitzades. P. ex., en Python poden obtenir-se mitjançant la comanda `pip freeze > requirements.txt`

L'extracció de les dades s'ha realitzat amb Python de manera automàtica fent servir la llibreria Selenium. Tant aquesta com la resta de llibreries utilitzades es troben al fitxer **requirements.txt** del repositori on també hi ha inclòs el codi i el dataset a les carpetes: **source** i **dataset**, respectivament.

L'adreça del repositori és: <https://github.com/albert-queralto/scraping-energy-prices-spain>

A continuació, es detallen les accions que duu a terme el codi:

En primer lloc, s'inicialitza el WebDriver que utilitza un **User-Agent** aleatori cada vegada que es connecta per evitar ser identificats com un bot. La pàgina inicial que es carrega és <https://www.esios.ree.es/es>. Des d'allà, el bot navega de manera autònoma a la pàgina **Mercados y Precios** fent clic a l'enllaç de la pàgina principal.

Un cop a la pàgina de **Mercados y Precios**, la navegació per les dates escollides es podia realitzar fent clic als elements de la pàgina web identificats en la **Fig 5**.

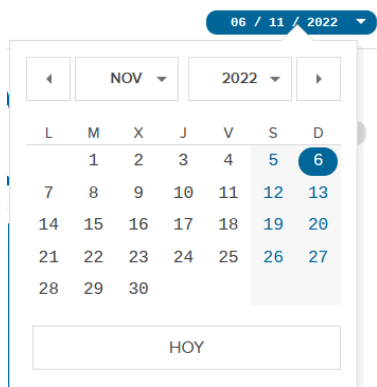



Fig 5. Exemple del calendari per navegar a diferents dates.



Fig 6. Exemple del selector horari de la pàgina **Mercados y precios**.

Tot i que la navegació era senzilla a l'hora de seleccionar el mes i l'any, els dies estan ubicats en una taula que, a més, classifica en divisors separats els dies que formen part del cap de setmana. Per això, per a realitzar aquest pas específic s'ha optat per utilitzar l'adreça web per a la navegació (<https://www.esios.ree.es/es/mercados-y-precios?date=01-11-2022>), és a dir, modificar els elements que corresponen al dia, mes i any de l'adreça web.

Un altre punt important és la selecció de l'hora. En aquest cas, es pot veure com el valor de l'hora es troba amagat en dos sub-menús. Per poder-hi accedir, s'ha fet clic als diferents elements que s'han localitzat amb el valor XPATH corresponent. Algunes dificultats trobades a l'hora de fer la selecció horària i les seves solucions s'exposen a continuació:

- Les classes dels divisors del codi HTML canvien dinàmicament en fer click i, per tant, no apareixen si abans no s'ha clicat a l'element ja que estan amagades. Per tant, s'ha hagut de comprovar el valor de les classes abans i després de fer clic a l'element per continuar amb el següent pas. En cas contrari, es torna a repetir el pas.
- La selecció del valor d'hora correcta fent servir la localització amb el webdriver no sempre és possible ja que queda amagat en la llista desplegable i no s'ha trobat l'element per desplaçar cap abaix la llista. Això retorna un error que s'ha resolt utilitzant la tecla  que desplaça la finestra cap a baix i permet trobar l'hora correcta. A més, ens hem assegurat que el valor de l'hora seleccionada coincideix amb l'hora correcta. En cas contrari, es torna a fer la cerca del valor.
- La resolució al problema anterior fa que a partir d'un cert nombre de desplaçaments de la finestra amb el teclat, l'element de selecció de les hores no sigui visible i, per tant, aparegui un altre error. Per tal de solucionar aquest inconvenient, s'ha procedit a fer un **scroll** de la finestra cap a la posició inicial amb **javascript**.
- Per algun motiu desconegut, la selecció d'algunes hores concretes no era possible en alguns dies. Per exemple, pel 27-03-2022 no es podia seleccionar les 02:00, tot i provar de fer la selecció manualment per descartar que fós un problema amb el bot. Per això, s'ha fet una implementació per reintentar un cert nombre de vegades d'agafar l'hora i en cas què es sobrepassi aquest límit, continuar amb la següent hora del dia.

Un cop ens trobem en el dia i hora indicats, s'obtenen les dades exemplificades a **Fig 2**. Les dades de la pàgina **Mercados y Precios** es guarden per dia en fitxers .csv independents per evitar pèrdues d'informació en cas que el bot aturi l'extracció.

Finalitzada l'extracció de les dades de **Mercados y Precios**, el bot navega a la pàgina **Generación y Consumo** de manera equivalent a la pàgina **Mercados y Precios**, és a dir, clicant a l'enllaç de la pàgina principal. Un cop allà, hem tornat a utilitzar el mateix mètode anterior per a navegar per les dates utilitzant l'adreça web (<https://www.esios.ree.es/es/generacion-y-consumo?date=01-11-2022>). El selector horari de

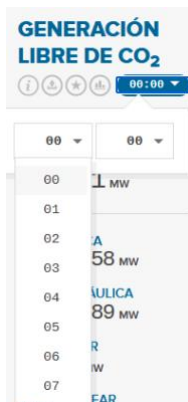


Fig 7. Exemple del selector horari de la pàgina **Generación y Consumo**.

la pàgina **Generación y Consumo** conté dos camps corresponents a les hores i els minuts. Per tant, s'ha hagut de trobar l'element correcte que permetia seleccionar el menú desplegable de l'esquerra fent servir l'identificador XPATH. Aquí, també hem trobat els mateixos problemes esmentats per la pàgina **Mercados y Precios** que s'han resolt de la mateixa manera.

Un cop seleccionada l'hora correcta, s'han obtingut les dades exemplificades a **Fig 3**. Les dades de la generació d'energia es formaten adequadament per crear la columna **Energy Source** que conté el nom de la font d'energia i **Energy Generation** amb el valor dels MW generats. Per tal d'emmagatzemar les dades, el bot agafa els fitxers amb les dades de **Mercados y Precios** (fitxers per dia) i fa un **merge amb pandas** per dia i hora, tornant a guardar les dades en un nou fitxer.

Altres consideracions a tenir en compte:

- Per tal de simular el comportament d'un usuari i permetre al bot trobar els elements de la pàgina, s'han establert mecanismes d'espera que donaven prou temps a la pàgina a realitzar la càrrega dels elements dinàmics. Això s'ha fet de manera implícita amb **time.sleep** o en les opcions del webdriver.
- S'han gestionat diferents tipus d'excepcions com poden ser: **TimeoutException**, tornant a carregar la pàgina, o problemes amb els elements de la pàgina a l'hora de seleccionar les hores (**ElementClickInterceptedException**, **StaleElementReferenceException**, **ElementNotInteractableException** o **NoSuchElementException**), assegurant que les menús es mostren correctament i són visibles pel bot, tal i com s'ha explicat anteriorment.
- S'ha detectat que en alguns casos l'extracció de les dades no es guardava correctament, ja sigui perquè no es guardava el fitxer o perquè el fitxer guardat només contenia la capçalera de les columnes i cap dada. Per això, s'ha implementat una funció que compara les dates dels noms dels fitxers amb els valors presents en el rang definit inicialment. Si falta algun fitxer o no conté cap dada, es torna a executar l'extracció de dades pels fitxers faltants. Aquesta implementació s'ha fet tan per **Mercados y Precios** com per **Generación y Consumo**.
- En alguns casos l'execució del bot s'atura perquè els widgets no carreguen tot i haver deixat prou temps d'espera amb el WebDriver. Això podria ser un mecanisme establert pel propietari de la pàgina per prevenir el web scraping. L'única solució que s'ha trobat és recarregar la pàgina web fins que respongui.

Finalment, totes les dades es combinen i es guarden en un únic fitxer que conforma el dataset entregat.

Algunes inconsistències que presenta el dataset final es mencionen a continuació:

- Impossibilitat de descàrrega d'algunes hores de la pàgina web (com per exemple les 2:00 pel dia 27 de març de 2022). Aquest manca de dades es produeix degut a un problema en l'extracció de les dades, probablement degut a què el bot no detecta l'element correcte identificat amb l'XPath.
- La descàrrega les dades es fa en format text, per la qual cosa caldrà adequar el format al tipus de columna abans d'utilitzar-se per a un posterior anàlisi.
- En algunes columnes, el valor numèric va acompanyat de les unitats com, per exemple, les columnes **renewable generation (%)** i **renewable generation (MW)**. Per tant, farà falta realitzar les transformacions pertinents per obtenir únicament els valors numèrics.
- Finalment, la columna **energy_source** conté el nom de les columnes originals i caldria realitzar algunes transformacions perquè representi millor la informació continguda.

10. Dataset. Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset (<https://doi.org/...>). El dataset també haurà d'incloure's a la carpeta **/dataset** del repositori.

Si existeix qualsevol circumstància que impedeixi publicar obertament el dataset real a Zenodo, s'haurà de: (1) comentar aquesta circumstància i justificar el motiu en aquest apartat; (2) generar un dataset simulat i publicar-lo a Zenodo, obtenint l'enllaç del DOI; i (3) comunicar al professor el dataset real de manera privada (p. ex., utilitzant un repositori privat).

El dataset en format CSV es troba dins la carpeta **dataset** a Github, el qual es pot consultar en el següent [enllaç](#). Tanmateix, es pot trobar publicat a Zenodo accedint mitjançant la següent [adreça](#) (DOI: 10.5281/zenodo.7335618).

11. Vídeo. Realitzar un breu vídeo explicatiu de la pràctica (**màxim 10 minuts**), que haurà de comptar amb la participació dels dos integrants del grup. Al vídeo s'haurà de realitzar una presentació del projecte, destacant els punts més rellevants, tant de les respostes als apartats com del codi utilitzat per a extreure les dades. Indicar l'enllaç del vídeo (<https://drive.google.com/...>), que haurà d'estar al Google Drive de la UOC.

El vídeo es pot trobar en el següent [enllaç](#) de Google Drive, al qual ja s'ha donat accés a la professora.

12. Taula de contribucions. Al final del document, ha d'aparèixer la següent taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació per part del grup que l'integrant ha participat a aquest apartat.

Contribucions	Signatura
Investigació prèvia	AQ, EM
Redacció de les respostes	AQ, EM
Desenvolupament del codi	AQ, EM
Participació al vídeo	AQ, EM

Referències:

[1] Alexis Romero, "La guerra a Ucraïna i la creixent inflació ressusciten el debat sobre la intervenció del mercat elèctric" (Público, 01/03/2022). <https://www.publico.es/public/encariment-preus-guerra-ucraina-i-creixent-inflacio-ressusciten-debat-intervencio-mercat-electric.html>

[2] Vicenç Batalla, "La UE frenarà la pujada de l'electricitat per la guerra" (El Punt Avui, 12/03/2022). <https://www.elpuntavui.cat/politica/article/17-politica/2111814-la-ue-frenara-la-pujada-de-l-electricitat-per-la-guerra.html>

[3] "El món a RAC1", "La inflació, pels núvols: "Ens empobrirem com quan vam canviar la pesseta per l'euro" (RAC1, 01/11/2022). <https://www.rac1.cat/el-mon/20221101/101928/inflacio-disparada-encariment-preus-energia-lloguers-hipoteques.html>

[4] Antonio Barrero, "El precio de la luz en España, un 35% por debajo de lo que pagan los franceses" (Renewable Energy Magazine, 28/09/2022). <https://www.energias-renovables.com/panorama/el-precio-de-la-luz-en-espana-20220928>

[5] OMIE, “Precio de la factura de la luz, datos y estadísticas” (epdata, 04/11/2022).

<https://www.epdata.es/datos/precio-factura-luz-datos-estadisticas/594>

[6] OMIE, MIBGAS i SendeCO2, “Evolución del precio de la luz, el gas y el CO₂” (epdata, 18/05/2022).

<https://www.epdata.es/datos/precio-factura-luz-datos-estadisticas/594>

[7] REE, “Demanda y generación de energía eléctrica en España, en gráficos” (epdata, 03/11/2022).

<https://www.epdata.es/datos/demanda-generacion-energia-electrica-espana-datos-graficos/636>