# Tipologia i cicle de vida de les dades
# Pràctica 1 - Web Scraping

Joel Rosell Mirmi
Víctor Iruela Garrido

## Apartat 1: Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació:

La informació recolectada es basa en el preu dels futurs elèctrics financers, els quals poden anar variant en funció del mercat. El lloc web del qual es recolecta la informació és el del MEFF (*"Mercado Oficial de Opciones Y Futuros Financieros en España"*), que disposa d'una API per a consultar les dades dels futurs dels preus de l'electricitat, a temps real. Aquesta API, però, té un cost i, en un primer moment, pot ser útil disposar de les dades que el lloc web publica diàriament de forma gratuïta, amb l'estructura que ens interessi. D'aquesta manera, es poden dur a terme processos analítics sobre elles i aconseguir informació rellevant amb un cost relativament baix.

El MEFF és l'entitat que gestiona els preus dels futurs elèctrics, per tant, tenen una certa obligació a publicar aquestes dades, per tal d'informar de com van variant aquests preus.

## Apartat 2: Definir un títol pel dataset. Triar un títol que sigui descriptiu:

Títol del dataset: "*data_inici_període-data_fi_període*_dfFutures.csv"

## Apartat 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 triat):

El dataset que hem escollit és un conjunt de dades que es correspon amb els preus de mercat dels futurs elèctrics. Les dades estan compostes per quatre variables o atributs: *'Dia'*, *'Preu Base (€)'* i *'Preu Pic (€)'* del futur elèctric. S'ha pres la variable *'Dia'* com a índex del dataset.

## Apartat 4: Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment:

## Apartat 5: Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit:

El format de les columnes del set de dades és el següent:

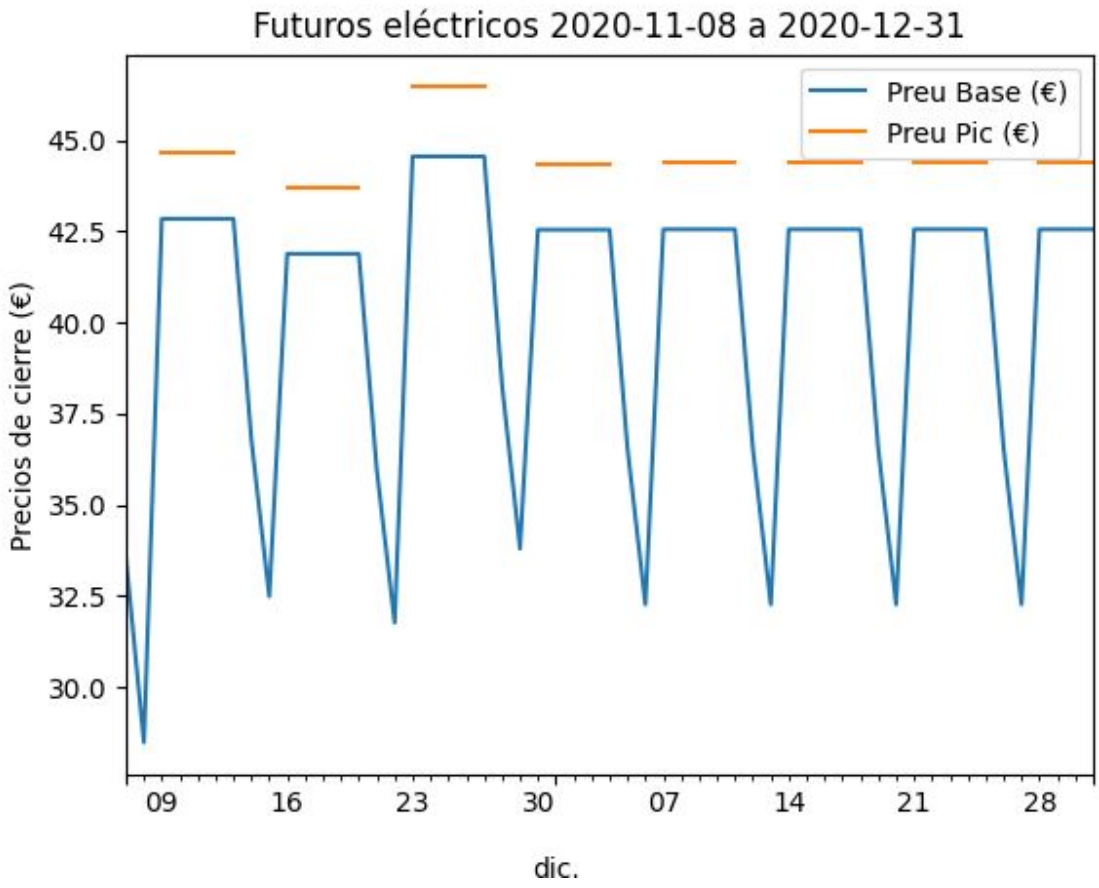
| Dia | Preu Base (€) | Preu Punta (€) |
|---|---|---|
| Datetime (yyyy-mm-dd) | float | float |

La columna *'Dia'* té cada una de les dates dels registres.
La columna *'Preu Base'* té el preu base dels futurs financers.
La columna *'Preu Punta'* té el preu punta dels futurs financers.

Les dades recullen els futurs a dos mesos vista (totes les disponibles al web), des de la data a la qual s'executi el programa. Els preus dels futurs són els que tenen aquell dia (cada dia s'actualitzen i poden canviar).

Les dades s'han recollit fent servir un scraper en *Python*. L'scraper utilitza el navegador integrat a l'eina *Selenium*, donant-li instruccions perquè vagi navegant fins a arribar a l'URL on són les dades. Una vegada allà, l'scraper busca la taula de les dades i l'emmagatzema en una variable en format d'string, perquè sigui més fàcil de parcejar. Després, parceja el text separant la informació en les tres variables descrites i, finalment, les guarda en un arxiu en format CSV.

Representació gràfica del conjunt de dades (futurs del dia 08/11/2020):



## Apartat 6: Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha):

El propietari de les dades és el “Mercado Oficial de Opciones Y Futuros Financieros en España” (**MEFF**).

## Apartat 7: Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre:

Aquestes dades són els preus de tancament dels futurs financers del mercat elèctric. Atès que si no és pagant, aquests mercats financers acostumen a ser molt hermètics, recollir aquestes dades pot ser molt interessant i útil. Podem crear una base de dades que emmagatzemi els preus diàriament i fer l’analítica posterior que ens interessi, sense haver de pagar l’API que ens proporcionen.

Aquestes dades podrien ser usades per una empresa que es dediqués a fer informes de consum elèctric, personalitzat per als clients de les comercialitzadores elèctriques. Aconseguir informació pública dels preus dels futurs elèctrics i adjuntar-la als informes, pot ser beneficiós per captar l'atenció dels usuaris més avançats. A més, aconseguir informació a temps real dels preus dels futurs és molt costós.

En un àmbit més analític, les dades també es podrien utilitzar per a analitzar els canvis en els preus dels futurs, en funció del dia de consulta (ja que les dades s’actualitzen diàriament). Així es podrien detectar, per exemple, dies de la setmana on els futurs baixen de preu.

## Apartat 8: Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

Llicència seleccionada:

○ **Released Under CC0: Public Domain License**

## Apartat 9: Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R:

Codi Python adjuntat, juntament amb aquest pdf.

## Apartat 10: Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció:

Dataset publicat a Zenodo, en format CSV.

https://zenodo.org/record/4263314#.X6hGlnVKhH4

## Apartat 11: Quadre de contribucions:

| Contribucions | Signa |
|---|---|
| **Recerca prèvia** | JRM, VIG |
| **Redacció de les respostes** | JRM, VIG |
| **Desenvolupament del codi** | JRM, VIG |