Pràctica 1: Web scrapping

NOM I COGNOMS: Jordi Gual Obradors | Daniel Lijia Hu

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

El conjunt de dades es correspon amb els preus del mercat de metalls utilitzats com a matèries primeres, en concret de l'alumini, coure, or i plata. Aquests metalls poden tenir relacions en els cicles d'alces i baixes entre elles, i volem observar si alguns van avançats o retardats.

Les fonts utilitzades són dues:

- Per a les dades diàries, el lloc web és Yahoo Finance, que s'actualitza constantment amb els últims valors de la borsa pel que fa a aquests materials.
- Per a les dades històriques, el lloc web és IndexMundi que proporciona sèries històriques de molts béns de diferent tipus.

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

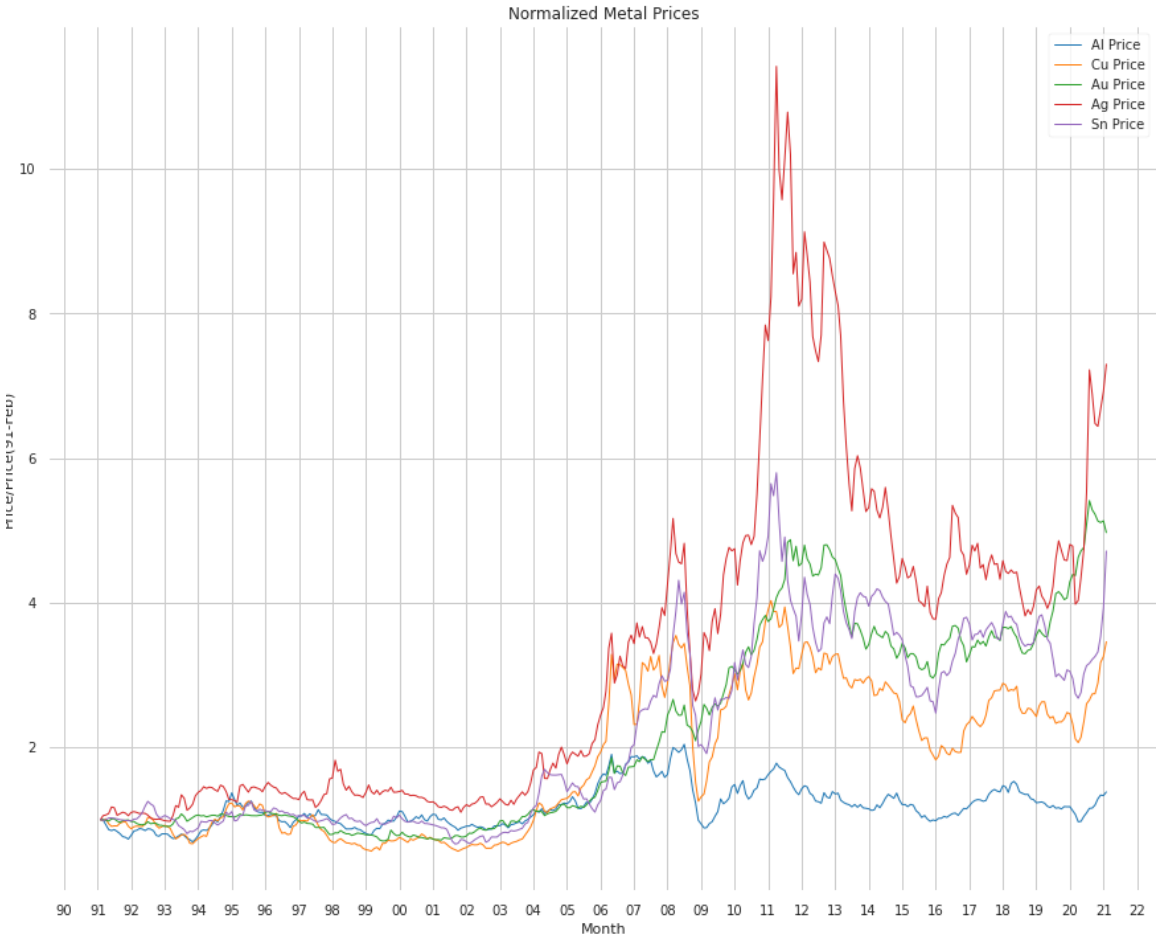
Evolució dels preus de diversos metalls

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).

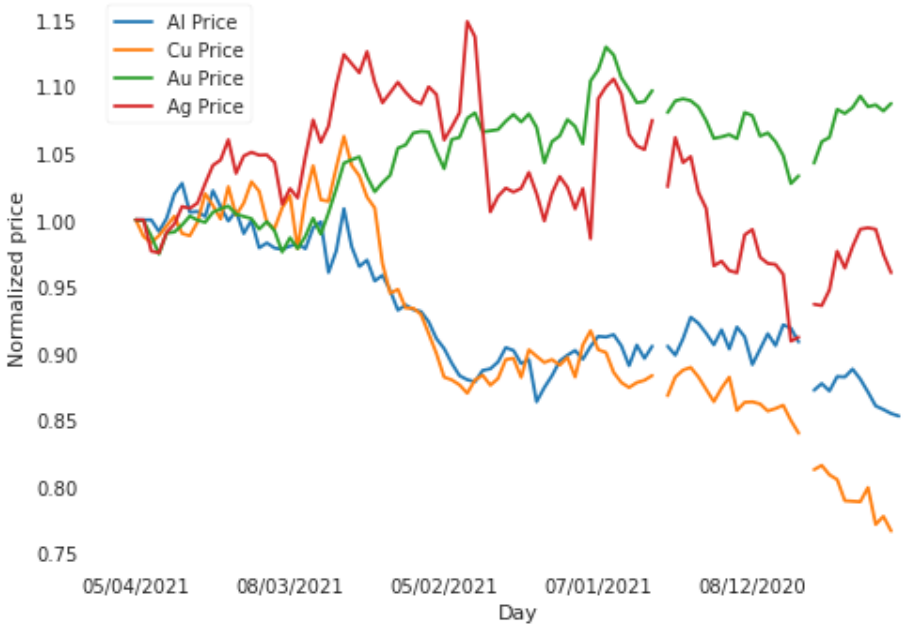
Preus internacionals en USD de l'alumini, coure, or i plata, històrics en base mensual i diaris dels últims 100 dies

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

Dades històriques



Dades scrappades



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

Per cada dia dels últims 100 dies, s'obté un registre en el conjunt de dades on es recullen les següents informacions:

- **Day:** Dia amb el format dd/mm/aa
- **Al Price:** Preu de l'alumini a l'inici del dia en USD
- **Cu Price:** Preu del coure a l'inici del dia en USD
- **Au Price:** Preu de l'or a l'inici del dia en USD
- **Ag Price:** Preu de la plata a l'inici del dia en USD

Per la part històrica, preus mitjans mensuals dels últims 30 anys.

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars.

Les dades més adients per a la nostra anàlisi serien les que proporciona sota pagament i completament 'inescrappables' la borsa de metalls internacional més important, la London Metal Exchange (LME). LME ofereix les cotitzacions i volums de contractació diaris de períodes superiors als 30 anys, que finalment hem considerat els adequats. Així, amb una granularitat diària, es poden buscar interrelacions de resposta molt ràpida (aplicacions de trading en el mateix dia) i de resposta setmanal-mensual fent agrupacions.

Com alternativa, hem procurat construir un dataset gratuit adient per a l'objectiu, a partir de dues fonts: Una amb dades estàtiques i l'altra d'actualització diària via scrapping.

La font del dataset estàtic és IndexMundi
https://www.indexmundi.com/es/precios-de-mercado/

IndexMundi és un portal de dades que recull fets i estadístiques provinents de múltiples fonts i les transforma en gràfics fàcils d'utilitzar per a una audiència global. Recull estadístiques que es troben disperses o no fàcilment visibles i les presenta en forma de mapes, gràfics, i taules entenedores amb un cop d'ull.

Els seus fundadors són Miguel Barrientos i Claudia Soria, amb experiència en els camps de Business Intelligence i Data Warehouses el primer, i en economía i en emprenedoria la segona.

D'aquesta web hem obtingut els preus mitjans mensuals dels últims 30 anys per als metalls: Alumini, Coure, Plata i Or. Considerem que 30 anys aporten un nombre suficient de dades per a poder veure tendències a velocitats moderades o lentes.

Per altra banda, de cara a complementar la utilitat del dataset, hem buscat fer el scrapping de les dades diàries fornides per Yahoo, amb l'objectiu de mantenir les dades mensuals, i anar construint un dataset gratuit de granularitat diària per a l'anàlisi de variacions ràpides.

Com cites d'anàlisi semblants en podem trobar una multitud:

- Macrotrends: https://www.macrotrends.net/charts/precious-metals.
- Trading economics: https://tradingeconomics.com/forecast/commodity
- Banc Central Europeu: https://www.ecb.europa.eu/pub/pdf/other/ebbox201708_01.en.pdf
- MetalMiner: https://agmetalminer.com/monthly-report-price-index-trends/

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.

L'interès per al públic en general del dataset és força evident: Aquesta informació pot servir de base per a les decisions de trading d'aquests metalls. De fet, a part de la LME hi ha múltiples webs d'inversió on, també amb pagament previ, es poden aconseguir les dades.

Ja més des d'un punt de vista particular, la motivació per a la preparació d'aquest dataset ha estat doble:

- Per una banda buscar un primer contacte en la recerca de dades de mercat per a un possible treball més complet en sèries temporals. Malauradament les eines adequades per a fer anàlisis profundes encara no les dominem i ens quedarem en la part EDA.
- Per altra banda, centrar-se en un mateix mercat però amb presència de productes que tradicionalment han estat considerats de comportaments diferenciats (commodities vs valors refugi: Coure, Alumini vs Or o Plata)

Com en tot treball de recerca en dades, a mida que aconsegueixes respostes es generen noves preguntes. Però les preguntes inicials i més immediates són:

- Quines correlacions existeixen entre els preus dels diferents metalls
- A quines escales temporals trobem les correlacions.
- Presència de predictors entre els diferents preus.

Aquest dataset permet respondre aquestes preguntes i algunes més que segur sorgiran en el procés de mineria.

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

La llicència escollida per la publicació d'aquest conjunt de dades ha estat CC BY-NC-SA 4.0 License. Seguidament citem els motius pels quals hem elegit aquesta llicència:

- **Attribution:** S'ha d'acreditar pròpiament, proveir d'un enllaç a la llicència i indicar si s'ha fet algun canvi. Així es reconeix la nostra feina i veure realment quina informació extra s'ha afegit de collita pròpia.

- **NonCommercial:** No es pot utilitzar el material per motius comercials. Considerem que la finalitat final de veure la dependència entre els metalls és altament beneficiosa per les empreses, i per tant bastant desitjable de treballar-hi amb l'objectiu de vendre la informació.

- **ShareAlike:** Les modificacions i addicions sobre el treball publicat sota aquesta llicència s'hauran de distribuir sobre la mateixa. Això fa que el treball segueixi els objectius i esència original dels autors.

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

El codi es divideix en dos diferents fitxers per dos etapes ben diferents:

- *MetalHistoricData.py*: serveix per representar gràficament les dades històriques del metall dels últims anys

- *MetalPriceScrapper.py*: serveix per obtenir les dades dels últims preus dels metalls en qüestió (Alumini, coure, or i plata), i crear-ne un .csv amb aquests.

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

El DOI que dóna accés al CSV en Zenodo és el següent:

http://doi.org/10.5281/zenodo.4667668

Contribucions	Signa
Recerca prèvia	jgualob, dhu
Redacció de les respostes	jgualob, dhu
Desenvolupament codi	jgualob, dhu