

## Tipologia i cicle de vida de les dades

### PRÀCTICA 1

Alumne: Francesc Albuera Reverte

#### [Apartat 1]

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

Després de llegir els apunts m'he dedicat força temps a buscar per internet llocs interessants per practicar web scraping i poder realitzar així la primera pràctica de l'assignatura.

No obstant, m'he trobat en que la majoria de pàgines que a priori em semblaven interessants, bé sigui per la temàtica o bé sigui per que a nivell de dades podria ser interessant explotar-les, a les polítiques d'ús i privacitat prohibeixen explícitament l'ús del web scraping. Per exemple, pàgines interessants a priori eren pcomponents o wikiloc.

Tal com anava buscant informació i també tutorials per internet sobre web scraping amb python, m'he topat amb la pàgina:

- <https://datosmacro.expansion.com/>

Aquesta pàgina no conté informació explícita que prohibeixi la utilització de tècniques de web scraping i per tant, es l'escollida per treballar la pràctica.

Aquesta pàgina podríem dir que aglutina dades de diverses fonts i les mostra en forma de llistats, gràfics, ...

Al navegar una mica per la pàgina, observo que d'entre tota la informació que conté, hi ha el deute que te cada país. Aquesta dada personalment la trobo interessant. A partir d'aquí, també he vist que en aquest espai web hi ha força informació de cada país.

Amb tot això, he decidit crear un dataset que contindrà les dades del deute de cada país, juntament amb unes certes dades més de cada país, dades que a priori són indicadors de diferents àmbits, i la idea és a posteriori explotar les dades per buscar relacions entre variables, per exemple, si existeix relació entre el salari mig i el deute, o entre el número d'assassinats i el deute, ...

Per tant, emmagatzemarem diversos indicadors de tots els països juntament amb el deute de cadascun d'ells per tal de buscar relacions entre variables.

#### [Apartat 2]

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

El dataset podria portar per títol:

- *Deute per país juntament amb indicadors rellevants*

### [Apartat 3]

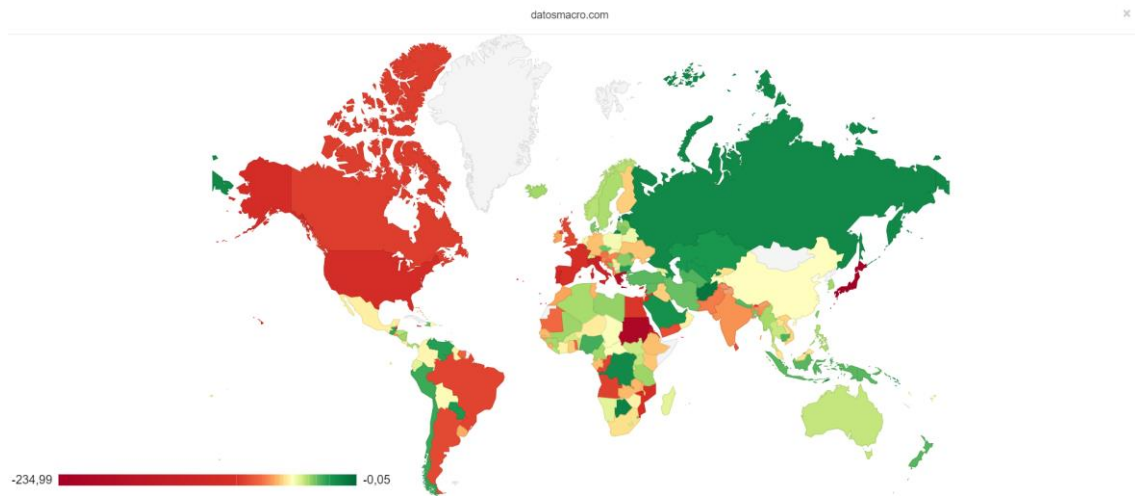
**Descripció del dataset.** Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

Llistat dels països del món on per cada país tenim el deute en diferents formats, i tenim diversos indicadors del país en qüestió susceptibles d'establir relacions.

### [Apartat 4]

**Representació gràfica.** Presentar una imatge o esquema que identifiqui el dataset visualment

Deute al món:



### [Apartat 5]

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

El dataset generat, cada columna té el seu 'header' i els llistarem a continuació:

- **Países:** nom de cada país.
- **Fecha:** anualitat del deute informat.
- **Deuda total (M.€):** deute total en milions d'euros.
- **Deuda (%PIB):** deute corresponent al % del PIB.
- **Deuda Per Cápita:** deute per càpita.
- **Anualidad índice corrupción:** anualitat de l'obtenció de l'índex de corrupció.
- **Índice de corrupción:** índex que classifica al països puntuant de 0 ( percepció d'un nivell alt de corrupció ) a 100 ( percepció d'un nivell baix de corrupció ) en funció de la percepció de corrupció del sector públic que tenen els seus habitants.
- **Anualidad índice fragilidad:** anualitat de l'obtenció de l'índex de fragilitat.
- **Índice mundial de fragilidad:** índex que va de 0 a 100 i que es calcula en funció de si un país té una sèrie de característiques com poden ser pèrdues de terrenys, conflictes bèl·lics, incapacitat de proporcionar serveis públics bàsics, incapacitat per interaccionar amb altres països d'una comunitat internacional, ...
- **Fecha tasa desempleo:** anualitat de l'obtenció de la taxa d'atur.
- **Tasa de desempleo:** percentatge de persones que estan a l'atur.
- **Fecha parados:** anualitat de l'obtenció del nombre d'aturats.
- **Parados:** nombre total de persones que estan en situació d'atur.

- **Anualidad salario mínimo:** anualitat de l'obtenció del valor del salari mínim.
- **Salario mínimo:** import salari mínim interprofessional en euros.
- **Anualidad salario medio:** anualitat de l'obtenció del valor del salari mig.
- **Salario medio:** import de la mitja del salari en euros.
- **Anualidad población:** anualitat de l'obtenció del numero total de població.
- **Población:** numero total d'habitants.
- **Anualidad inmigrantes:** anualitat de l'obtenció del numero total d'immigrants.
- **Inmigrantes:** numero total d'immigrants.
- **Anualidad emigrantes:** anualitat de l'obtenció del numero total d'emigrants.
- **Emigrantes:** numero total d'emigrants.
- **Anualidad homicidios intencionados:** anualitat de l'obtenció del numero d'assassinats.
- **Homicidios intencionados:** numero total d'assassinats intencionats.
- **Anualidad esperanza de vida:** anualitat de l'obtenció de l'esperança de vida.
- **Esperanza de vida:** edat que defineix l'esperança de vida.
- **Anualidad emisiones CO2 toneladas per capita:** anualitat de l'obtenció de l'índex de les emissions de CO2 toneladas per càpita.
- **Emisiones CO2 toneladas per capita:** numero de toneladas de CO2 emeses a l'atmosfera.

Podem veure que diversos indicadors es basen en una anualitat concreta. Per altra banda, la manera com hem obtingut els diferents camps ho expliquem a l'apartat 9 i al propi notebook on esta el codi implementat.

#### [Apartat 6]

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

Les dades s'han extret fent scraping de la web <https://datosmacro.expansion.com/>.

#### [Apartat 7]

**Inspiració.** Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

La idea de generar aquest dataset es per aplicar models de mineria de dades per mirar d'extreure relacions entre el deute d'un país i diferents indicadors sobre el país en qüestió.

L'objectiu es poder descobrir indicadors que influeixen amb el deute d'un país per a partir del seu descobriment, investigar sobre aquest indicador per aportar millores.

#### [Apartat 8]

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

Aquesta pràctica s'ha basat en agafar dades d'una pàgina que en si mateixa agafa les dades dels portals que es llisten en aquest enllaç:

<https://datosmacro.expansion.com/legal/fuentes>

Per tant, estem consumint dades d'un portal que consumeix dades de portals fets per a tal fi. Per tant, d'alguna manera estem generant un dataset concret per a una finalitat que podria ser analitzada per tercers, amb fins d'estudi universitari per exemple.

Amb tot això, la llicència podria ser Released Under CC0: Public Domain License, és a dir, sense drets de propietat intel·lectual. Hem generat un dataset per a compartir lliurement amb la comunitat.

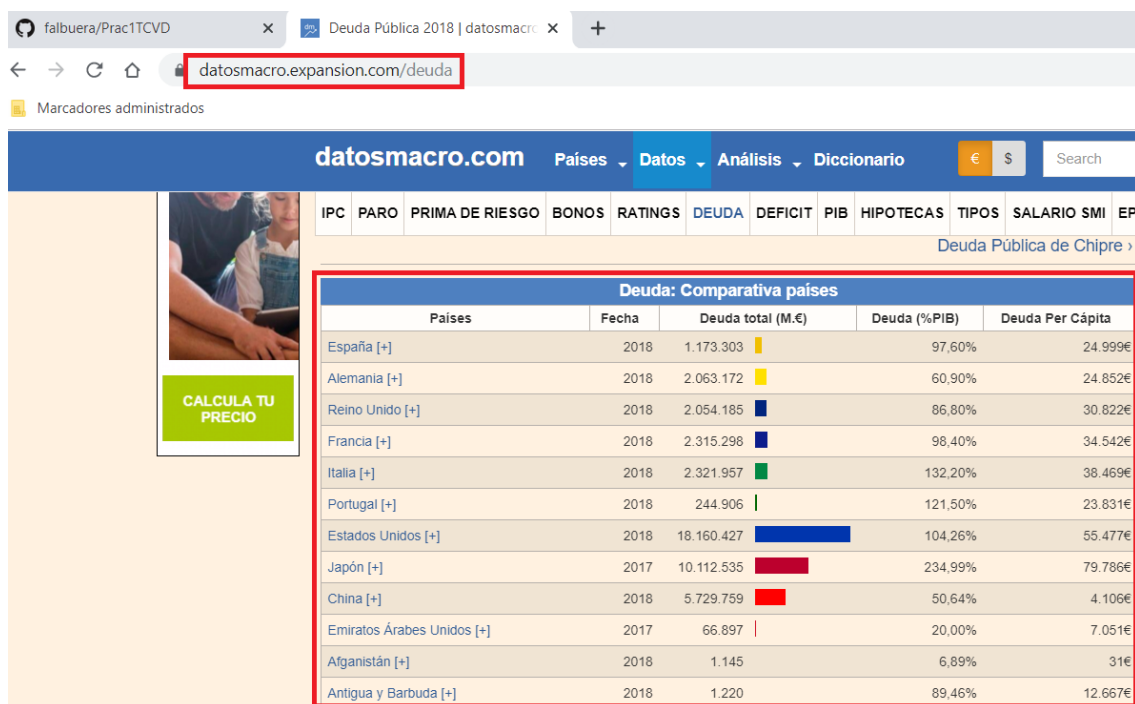
### [Apartat 9]

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

Al repositori github esta pujat el fitxer **francesc\_albuera\_prac1.ipynb**. Al propi notebook estan els comentaris que expliquen el codi en python.

En tot cas, aquí explicarem quina era la informació que volíem extreure i quins problemes se'ns plantejaven.

La idea era de la pàgina <https://datosmacro.expansion.com/deuda> :



Deuda: Comparativa países					
Países	Fecha	Deuda total (M.€)	Deuda (%PIB)	Deuda Per Cápita	
España [+]	2018	1.173.303	97,60%	24.999€	
Alemania [+]	2018	2.063.172	60,90%	24.852€	
Reino Unido [+]	2018	2.054.185	86,80%	30.822€	
Francia [+]	2018	2.315.298	98,40%	34.542€	
Italia [+]	2018	2.321.957	132,20%	38.469€	
Portugal [+]	2018	244.906	121,50%	23.831€	
Estados Unidos [+]	2018	18.160.427	104,26%	55.477€	
Japón [+]	2017	10.112.535	234,99%	79.786€	
China [+]	2018	5.729.759	50,64%	4.106€	
Emiratos Árabes Unidos [+]	2017	66.897	20,00%	7.051€	
Afganistán [+]	2018	1.145	6,89%	31€	
Antigua y Barbuda [+]	2018	1.220	89,46%	12.667€	

Agafar les dades de tots els països de la taula. Aquí haviem de tenir en compte que:

A la columna deute total, teníem que anar en compte ja que:

Deuda pública de Chile

```
jQuery("#tbi_318").tablesorter();
);
</script>
<table id="tbi_318" class="table tabledat table-striped table-condensed table-hover">
  <thead></thead>
  <tbody>
    <tr>
      <td></td>
      <td class="fecha" data-value="2018-12-01">2018</td>
      <td class="numero eun" data-value="1173303">1.173.303</td>
      <td class="hbar.eur.wdsp1.2"></td> == $
      <td class="numero dol" data-value="1385671">1.385.671</td>
      <td class="modal.doi.wdsp1.2"></td>
      <td class="numero" data-value="97,6">97,60%</td>
      <td class="numero eun" data-value="24999">24.999€</td>
      <td class="numero dol" data-value="29523">
```

Deuda: Comparación

td.hbar.eur.wdsp1.2 89.6 × 26


Países	Fecha	Deuda Total	Deuda Per Cápita
España [+]	2018	1.173.303	97,60% 24.999€
Alemania [+]	2018	2.653.172	60,90% 24.852€
Reino Unido [+]	2018	2.051.185	86,80% 30.822€
Francia [+]	2018	2.315.298	98,40% 34.542€
Italia [+]	2018	2.321.957	132,20% 38.469€
Portugal [+]	2018	244.906	121,50% 23.831€
Estados Unidos [+]	2018	18.160.427	104,26% 55.477€
Japón [+]	2017	10.112.535	234,99% 79.786€
China [+]	2018	5.729.759	50,64% 4.106€

Aquesta columna internament té una columna que mostra un gràfic i també una altra columna que conté l'import en dòlars. Nosaltres ens quedarem sempre amb l'import en euros ( això al codi està comentat com s'ha fet ).

Després també s'aprofita l'enllaç que hi ha a la primera columna de cada fila per anar al detall de cada país per poder extreure més informació:

Deuda: Comparativa países				
Países	Fecha	Deuda total (M.€)	Deuda (%PIB)	Deuda Per Cápita
España [+]	2018	1.173.303	97.60%	24.990€
Alemania [+]	2018	2.063.172	60.90%	24.852€
Reino Unido [+]	2018	2.054.105	86.80%	30.822€
Francia [+]	2018	2.315.290	98.40%	34.542€
Italia [+]	2018	2.321.957	132.20%	38.469€
Portugal [+]	2018	244.906	121.50%	23.831€
Estados Unidos [+]	2018	18.160.427	104.26%	55.477€
Japón [+]	2017	10.112.535	234.99%	79.786€
China [+]	2018	5.729.759	50.64%	4.106€
Emiratos Árabes Unidos [+]	2017	66.897	20.00%	7.051€
Afganistán [+]	2018	1.145	6.89%	31€

Cada enllaç contenia ‘/deuda/espana’ i nosaltres necessitàvem anar a ‘/países/espana’ per anar a la pàgina que contenia més informació del país.

 x España: Economía y demografía x +

< > ↺ 🏠 🔒 **datasmacro.expansion.com/paises/espana**

📁 Marcadores administrados

---

**datasmacro.com** Países Datos Análisis Diccionario € \$ Search

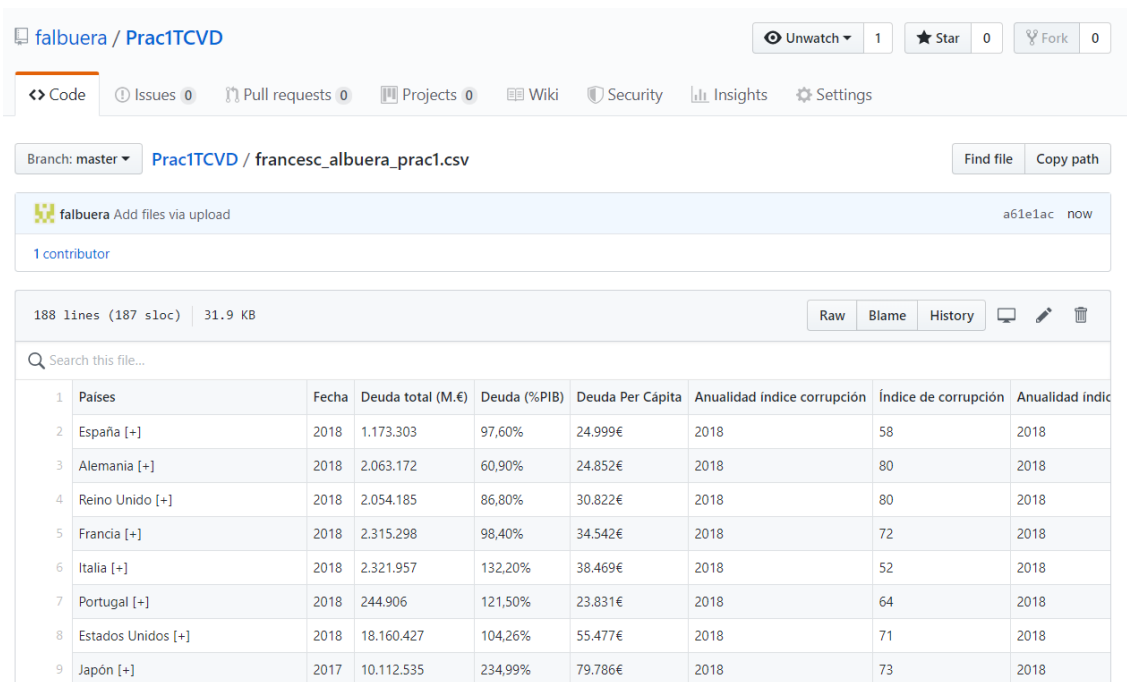
IPC	PARO	PRIMA DE RIESGO	BONOS	RATINGS	DEUDA	DEFICIT	PIB	HIPOTECAS	TIPOS	SALARIO SMI	EMI
[+]											
Gasto Defensa Per Capita [+]									2018		3284
Rating Moody's [+]									13/04/2018	Baa1	
Rating S&P [+]									20/09/2019	A	
Rating Fitch [+]									21/06/2019	A-	
Índice de Corrupción [+]									2018		51
Ranking de Competitividad [+]									2018		29
Índice de Fragilidad [-]									2018		41
Ranking de Trans. [+]									01/01/2019		87
Ranking de la Innovación [+]									2018		28
<b>Mercado Laboral</b>											
Tasa de desempleo [+]									Septiembre 2019		14,2%
Tasa de desempleo [+]									III Trim 2019		13,9%
Parados [+]									III Trim 2019		3.214 m
SMI [-]									2019		1.050,0 €
Salario Medio [+]									2018		26.923€
Ranking [-]									2017		44
<b>Mercados - Cotizaciones</b>											
Tipo de cambio del dólar [+]									08/11/2019		0.906€
Bono 10 años [+]									11/11/2019		0.40%
Prima Riesgo [+]									11/11/2019		61
Bolsa (Var. este año %)[+]									08/11/2019		10.00%

I el que s'ha fet ha estat aprofitar aquesta enllaços i fer un replace de 'deuda' per 'países' i d'aquesta manera teníem a cada fila de la taula l'enllaç per analitzar la pàgina de cada país i concatenar la informació desitjada.

## [Apartat 10]

### Dataset. Presentar el dataset en format CSV

Al repositori github esta pujat el fitxer **francesc\_albuera\_prac1.csv**. Adjuntem una captura de pantalla:



	Países	Fecha	Deuda total (M.€)	Deuda (%PIB)	Deuda Per Cápita	Anualidad índice corrupción	Índice de corrupción	Anualidad índice
1	España [+]	2018	1.173.303	97,60%	24.999€	2018	58	2018
2	Alemania [+]	2018	2.063.172	60,90%	24.852€	2018	80	2018
3	Reino Unido [+]	2018	2.054.185	86,80%	30.822€	2018	80	2018
4	Francia [+]	2018	2.315.298	98,40%	34.542€	2018	72	2018
5	Italia [+]	2018	2.321.957	132,20%	38.469€	2018	52	2018
6	Portugal [+]	2018	244.906	121,50%	23.831€	2018	64	2018
7	Estados Unidos [+]	2018	18.160.427	104,26%	55.477€	2018	71	2018
8	Japón [+]	2017	10.112.535	234,99%	79.786€	2018	73	2018

## Conclusions i millores

Un cop generat el dataset i finalitzada la pràctica, entenc que es podria fer una fase 2, on podríem netejar una mica les dades, ja que per exemple al nom dels països tenim els caràcters [+], tenim alguns camps buits, ... i per altra banda, si bé estem treballant amb països i aquests són finits i no són un numero molt elevat, el conjunt de dades obtingut és relativament petit. Pot ser es podria ampliar el dataset amb la informació d'anualitats anteriors de tots els indicadors que hem emmagatzemat per tal de tenir més volum de dades per a quan es vulgui aplicar un procés de mineria de dades.