Maria Begoña Felip
Vicenç Pio

Pràctica 1: Web scraping

1. Context.

Dins del període de pandèmia amb el que convivem fa més d'un any, ens hem decidit per buscar informació a través de ens públics on la informació sigui veraç.

Hem triat una web de codi obert (Dades Obertes) de la Generalitat de Catalunya. Disponible a Internet a: http://governobert.gencat.cat/ca/dades_obertes/.

La Generalitat de Catalunya ofereix un portal de dades d'accés públic, que permet pensar que és una bona opció per analitzar les dades obtingudes mitjançant Web Scraping.

Així doncs, l'objectiu concret d'aquesta activitat és la creació d'un dataset a partir de les dades obtingudes en la següent web disponible a Internet a: https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/jj6z-iyrp/data, extreta de la web arrel Govern Obert de la Generalitat de Catalunya i que conté el registre de casos de COVID-19 a Catalunya, segregats per sexe i municipi.

L'objectiu de la web és facilitar de forma pública i transparent el seguiment epidemiològic de la pandèmia de la COVID-19 mitjançant el monitoratge dels indicadors clau definits en el pla de control de la transmissió de la COVID-19 a Catalunya.

A diferència d'altres fonts, i per motius epidemiològics, s'imputa l'inici del cas en funció de la data més antiga que es disposa ja sigui del diagnòstic de COVID19 de l'Atenció Primària (si hi és) o de la realització de la PCR/TA+, ja que aquestes dues dates són les que indiquen l'inici del cas.

Aquestes dades provenen de diversos sistemes d'informació del Departament de Salut i del Servei Català de la Salut, detallats a https://dadescovid.cat/documentacio, i mostren per a cada dia, municipi, sexe i procediment diagnòstic el nombre de casos identificats com a positius COVID mitjançant alguna prova diagnòstica o per estudi epidemiològic. Tots els casos s'activen pels serveis de vigilància epidemiològica.

Les dades s'actualitzen diàriament. Només es mostren les dades fins 3 tres dies abans.

2. Definir un títol pel dataset.

CovidDataCatalonia.csv

3. Descripció del dataset.

La recollida de dades s'inicia el 01/03/2020 mitjançant diverses fonts com s'explica en l'apartat anterior. Conté:

- Files: 170K
- Columnes: 11

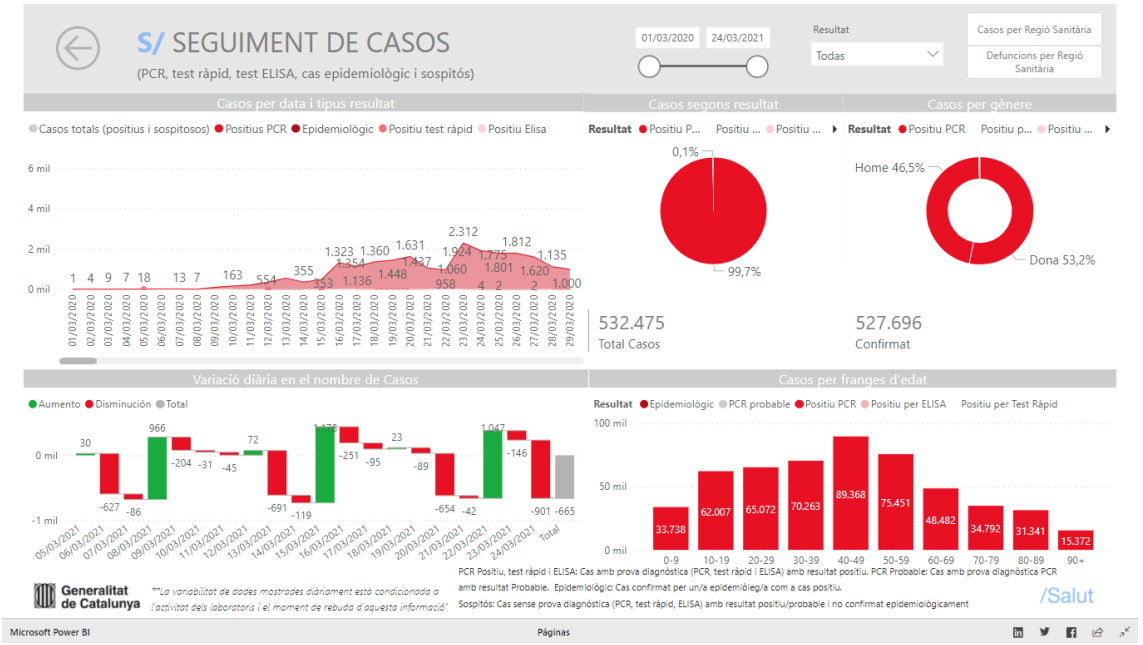
La informació del dataset prové de les dades del registre *RSAcovid19* del Departament de Salut, actualitzades diàriament, i inclou:

- Les dades dels **casos positius acumulats** que són aquells que han donat positiu en alguna prova diagnòstica (PCR o test ràpid).

- Les dades de **casos sospitosos acumulats** corresponen a persones que en algun moment han presentat símptomes i un professional sanitari els ha classificat com a possible cas, però no tenen una prova diagnòstica (PCR o test ràpid) amb resultat positiu.
- Tots ells són casos activats pel servei de vigilància epidemiològica.
- Les dades que es mostren són aquelles en les quals s'ha pogut identificar la zona de residència que consta a la targeta sanitària.

S'ha de tenir en compte que els casos de coronavirus SARS-CoV-2 estan subestimats perquè no s'hi inclouen les persones que contrauen la malaltia però tenen símptomes lleus i passen quasi desapercebuts o de forma subclínica. Aquesta dada queda recollida en la variable **TipuCasDescripcio** com a positiu o sospitós, comentada més endavant.

4. Representació gràfica.

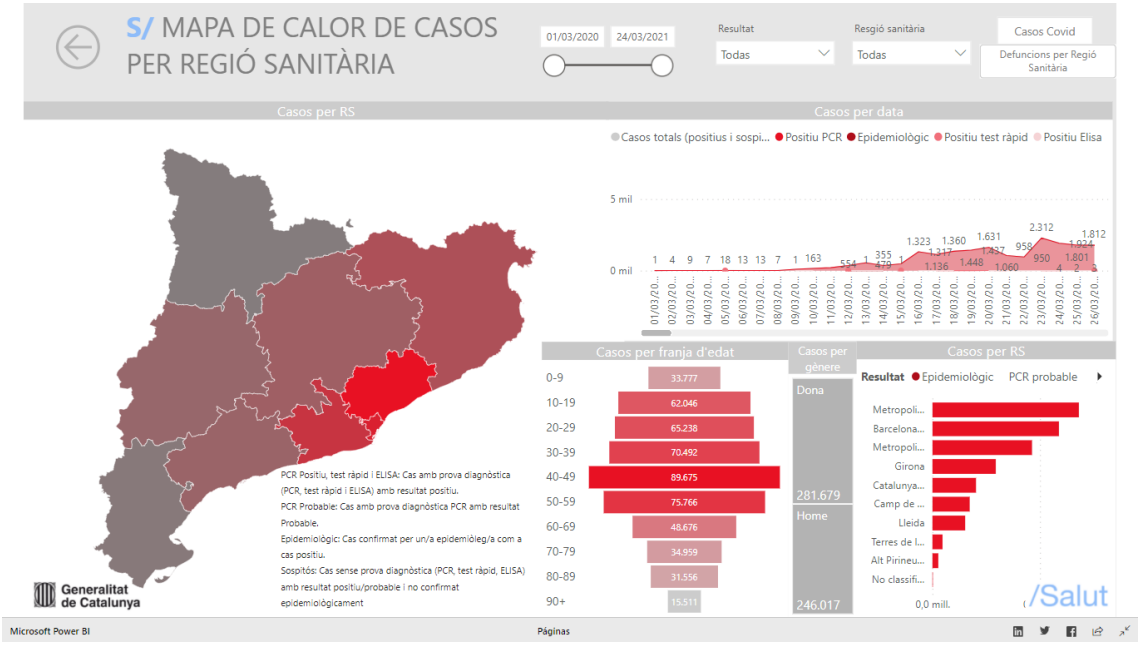


Disponible a Internet a:
https://app.powerbi.com/view?r=eyJrIjoiMGZlMDUzZDgtOWQ3MS00YTBhLWJjZjctYTJkNTg2NTRhOWQ4IiwidCI6IjNiOTQyN2RjLWQzMGUtNDNiYy04YzA2LWZmNzI1MzY3NmZlYyIsImMiOjh9

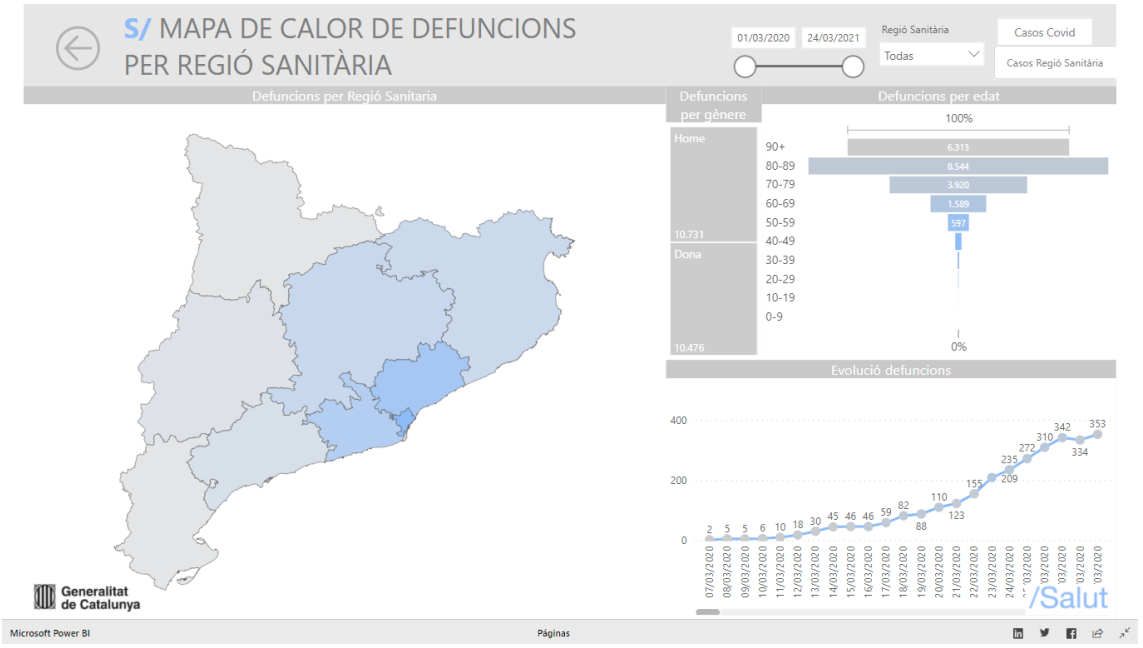
Representació geogràfica

La informació es representa d'acord a les divisions de regions, sectors i ABS sanitàries.

En els dos gràfics que venen a continuació, es representen les set regions sanitàries, delimitades a partir de factors geogràfics, socioeconòmics i demogràfics. Cada regió s'ordena, al seu torn, en sectors sanitaris, que agrupen les anomenades àrees bàsiques de salut, formades per barris o districtes a les àrees urbanes, o per un o més municipis en l'àmbit rural.



Disponible a Internet a:
https://app.powerbi.com/view?r=eyJrIjoiMGZlMDUzZDgtOWQ3MS00YTBhLWJjZjctYTJkNTg2NTRhOWQ4IiwidCI6IjNiOTQyN2RjLWQzMGUtNDNiYy04YzA2LWZmNzI1MzY3NmZlYyIsImMiOjh9&pageName=ReportSection5bb834a4f308a23a7fa0



Disponible a Internet a:
https://app.powerbi.com/view?r=eyJrIjoiMGZlMDUzZDgtOWQ3MS00YTBhLWJjZjctYTJkNTg2NTRhOWQ4IiwidCI6IjNiOTQyN2RjLWQzMGUtNDNiYy04YzA2LWZmNzI1MzY3NmZlYyIsImMiOjh9

5. Contingut.

El codi implementat per fer l'extracció del dataset està dividit en diferents parts:

- Funcions amb webdriver de la llibreria seleneca:
    - La funció *get_first_rows(nrows)* retorna les primeres n files del dataset.

    - La funció *get_column(n, column)* retorna els primers n registres de la columna seleccionada.

- Funció *download_web()* descarrega el contingut html.
- Funció *download_data(npages)* descarrega el dataset en format json. El mòdul json_to_csv.py permet crear un fitxer csv a partir del json anterior.

La base de dades té les següents columnes o camps:

a. **TipusCasData (data):** data de detecció del cas, format dd/mm/aaaa. La data del cas és la data d'inici de símptomes, no la data de realització de la prova diagnòstica. Correspon, doncs, a la data en que es va obtenir el resultat de la prova diagnòstica o en el cas sospitós la que la persona ha entrat en el registre *RSAcovid19*.

b. **ComarcaCodi (text):** codi de la comarca.

c. **ComarcaDescripcio (text):** Nom de la comarca.

d. **MunicipiCodi (text):** Codi del municipi. Aquesta informació correspon a les que consten a la targeta sanitària de la persona.

e. **MunicipiDescripcio (text):** Nom del municipi. En els casos en què no ha estat possible identificar el municipi de residència de la persona identificada com a cas positiu, el valor de la variable 'MunicipiDescripcio' és 'No classificat'. I en els casos de persones residents en municipis amb una població inferior a 200 habitants, per evitar la seva identificació i garantir-ne la confidencialitat, en aquesta variable s'hi fa constar el valor 'Altres municipis'.

f. **DistricteCodi (text):** variable per millorar el detall territorial de la informació facilitada per aquest conjunt de dades per a la població resident en el municipi de Barcelona. La identificació del districte s'obté a partir del sector sanitari, que per al municipi de Barcelona coincideix amb el districte municipal, i la codificació dels districtes que es mostra és la pròpia de l'Ajuntament de Barcelona. Per als registres corresponents a altres municipis, el camp 'DISTRICTECODI' està en blanc.

g. **DistricteDescripcio (text):** Nom del districte. Per als registres corresponents al municipi de Barcelona, aquests camps mostren els valors corresponents al districte municipal; per als registres corresponents a altres municipis conté el valor "No classificat".

h. **SexeCodi (text):** 0/1 (codi de sexe 0-home, 1-dona)

i. **SexeDescripcio (text):** descripció del sexe. Home/Dona o no classificat (no binari).

j. **TipusCasDescripcio (text):** especifica el procediment diagnòstic (epidemiològic, PCR probable, Positiu PCR, positiu per ELISA, positiu per Test Ràpid i positiu TAR)

k. **NumCasos (numèric):** variable nombre de casos. En aquesta variable es recull el nombre de casos acumulats segons la data en que es fa la prova diagnòstica en cas de resultat positiu. També poden entrar al registre per un sexe, un municipi o altres criteris especificats.

6. Agraïments.

Agraïm al **propietari** del lloc web les dades sobre les que estem treballant. Comencem, esbrinant qui és:

Amb la funció whois esbrinem el propietari del lloc web. Com es feu a la captura de pantalla, l'organització propietaria és la Generalitat de Catalunya que fa públiques les dades mitjançant el nom del domini transparenciacatalunya.cat, registrat com "Entorno Digital".

El portal de transparència es va crear en 2015 i està vigent i constantment actualitzat. Està sota els paràmetres de l'organització ICANN (Internet Corporation for Assigned Names and Numbers), la qual és una corporació sense ànims de lucre que coordina els identificadors únics dels ordinadors en tot el món. Ajuda a coordinar les funcions d'IANA (Internet Assigned Numberts Authority), que proporciona serveis clau crítics per a les operacions continues de Domain Name System o DNS.

El nom dels servidors són "dns.gencat.net" i "dns2.gencat.cat".

A més se'ns proporciona una adreça de correu electrònic per notificar mala praxis (abuse@entorno.es)

```
>>> print(whois.whois('https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/jj6z-iyrp/data'))
{
  "domain_name": "transparenciacatalunya.cat",
  "registrar": "Entorno Digital",
  "whois_server": null,
  "referral_url": null,
  "updated_date": "2021-02-27 23:49:11.115000",
  "creation_date": "2015-05-29 15:25:29.472000",
  "expiration_date": "2021-05-29 15:25:29.472000",
  "name_servers": [
    "dns.gencat.net",
    "dns2.gencat.cat"
  ],
  "status": "ok https://icann.org/epp#ok",
  "emails": "abuse@entorno.es",
  "dnssec": "unsigned",
  "name": null,
  "org": "Generalitat de Catalunya",
  "address": null,
  "city": null,
  "state": "BARCELONA",
  "zipcode": null,
  "country": "ES"
}
```

Com que les dades de la pandèmia mundial de COVID-19 es van començar a comptabilitzar i documentar des de març de l'any 2020, no hi ha gaire estudis diferents a d'altres períodes de temps. Són dades úniques davant d'una situació singular. Però sí que trobem anàlisis de dades obertes similars d'altres institucions o organitzacions, per exemple el de la **Universitat Politècnica de Catalunya** (https://biocomsc.upc.edu/en/covid-19/vaccination-cat) un anàlisis molt complert, però sobretot molt visual. Es tracta de mapes interactius de Catalunya amb les dades actuals i informació dels índex de risc respecte la epidèmia de Covid-19 per Municipi, Comarca, Àrea Bàsica de Salut i Regió Sanitària. Conté molta informació en format document i les dades mundials, també en forma de mapa interactiu.

Hi ha un altra anàlisis comparatiu que també resulta interessant, tot i que no es interactiu i les dades estan en forma gràfic, i és el que fa **betevé**, s'anomena Dades de la covid-19 a Catalunya: la corba, dia a dia (https://beteve.cat/societat/corba-coronavirus-catalunya/)

7. Inspiració.

L'interès de les dades està dins del context de la pandèmia que estem vivint des de fa més d'un any, centrant-nos en Catalunya, perquè és al territori on vivim. A més, que les dades estiguin constantment actualitzades a 3 dies (fins i tot l'estructura del conjunt de dades) i siguin públiques ha sigut un al·licient més.

Tot i que els dos estudis anteriors són molt interessants, ens hem decantat per la Generalitat de Catalunya perquè l'accés a les dades és obert i transparent. També ens hem trobat amb anàlisis gràfics, fàcils de copsar, igual que les dades.

Les preguntes que es pretenen respondre són les següents:

- Quina comarca ha registrat més casos positius?
- Hi ha més casos positius d'homes o de dones?
- En quines dates hi ha el major nombre de casos registrats? I el menor?
- Hi ha molta diferència entre el nombre de casos registrats a la ciutat de Barcelona i la resta de comarques?
- I entre els districtes a la ciutat, hi ha algun districte amb més casos que d'altres?

8. **Llicència**.

Hem seleccionat la llicència **Released Under CC BY-NC-SA 4.0 License** (Attribution-NonCommercial-ShareAlike 4.0 International) perquè volem compartir i enriquir la nostra feina amb d'altres companys i companyes, però que no se'n faci un ús comercial. Aquest llicència permet:

- **Compartir**: copiar i redistribuïr el material en qualsevol mitjà o format.
- **Adaptar**: remesclar, transformar i construir sobre el material.

Sota els termes següents:

- **Reconeixement:** s'ha de donar el crèdit adequat, proporcionar un enllaç a la llicència i indicar si s'han fet canvis. Es pot utilitzar de qualsevol manera, però en cap cas s'ha de suggerir que el propietari recolza l'ús que se'n faci.
- **No comercial:** no es pot utilitzar el material amb finalitats comercials.
- **Compartir igual:** si es remescla el material, es transforma o es fa servir, s'han de distribuir les contribucions sota la mateixa llicència que l'original.
- **No hi ha restriccions addicionals:** no es poden aplicar termes legals ni mesures tecnològiques que restringeixin legalment altres persones a fer allò que la llicència permet.

9. **Codi**.

Es troba a les carpetes de l'enllaç a Github: https://github.com/bfelip66/Practica1_952

10. **Dataset.**

El link del dataset a zenodo és el següent: https://zenodo.org/record/4663121

El DOI: 10.5281/zenodo.4663121

Contribucions	Signa
Recerca prèvia	Maria Begoña Felip, Vicenç Pio
Redacció de les respostes	Maria Begoña Felip, Vicenç Pio
Desenvolupament codi	Maria Begoña Felip, Vicenç Pio

Avaluació inicial

- **Arxiu robots.txt**:

Un arxiu robots. txt indica els rastrejadors dels cercadors quines pàgines o arxius del lloc web poden sol·licitar i quins no. Principalment, s'utilitza per a evitar que les sol·licituds que rep el lloc el sobrecarreguin (gestiona el trànsit dels rastrejadors); no és un mecanisme per a impedir que una pàgina web aparegui en Google.

L'arxiu robots consta d'un o més grups i per cada grup, d'una o diverses regles (cadascuna va en una línia diferent i distingeixen entre majúscules i minúscules), i cadascuna d'elles bloqueja o permet l'accés d'un determinat rastrejador a una ruta d'arxiu concreta d'un lloc web. En el nostre cas, l'arxiu robots especifica el següent:

Sitemap (instrucció opcional): Indica la ubicació del sitemap del lloc web que es vol rastrejar. L'URL ha de ser qualificada, ja que no es comproven alternatives (amb o sense www, o amb o sense http o https).

https://s3-eu-west-1.amazonaws.com/sa-socrata-sitemaps-eu-west-1-prod/sitemaps/sitemap-analisi.transparenciacatalunya.cat.xml

Nota: en totes les regles, menys en sitemap, s'utilitza el comodí "*" com a prefix, sufix o cadena de ruta. La "#" s'utilitza per afegir comentaris.

Grup 1:
User-agent: *
Crawl-delay: 1

Grup 2:
User-agent: gsa-crawler
Crawl-delay: 1

Grup 3:
User-agent: nys-crawler
Crawl-delay: 1

Grup 4:
User-agent: nys-qa-crawler
Crawl-delay: 1

Els **user-agent** de nom "gsa-crawler", "nys-crawler" i "nys-qa-crawler" no poden rastrejar els següents directoris, ni subdirectoris:

/browse?*&category=
/browse?*&federation_filter=
/browse?*&limitTo=
/browse?*&q=
/browse?*&sortBy=
/browse?*&tags=
/browse?*&view_type=
/browse/*?*&category=
/browse/*?*&federation_filter=
/browse/*?*&limitTo=
/browse/*?*&q=
/browse/*?*&sortBy=
/browse/*?*&tags=
/browse/*?*&view_type=
/*/browse?*&category=

/*/browse?*&federation_filter=
/*/browse?*&limitTo=
/*/browse?*&q=
/*/browse?*&sortBy=
/*/browse?*&tags=
/*/browse?*&view_type=
/page/*?*&category=
/page/*?*&federation_filter=
/page/*?*&limitTo=
/page/*?*&q=
/page/*?*&sortBy=
/page/*?*&tags=
/page/*?*&view_type=
/catalog/*?*&category=
/catalog/*?*&federation_filter=
/catalog/*?*&limitTo=
/catalog/*?*&q=
/catalog/*?*&sortBy=
/catalog/*?*&tags=
/catalog/*?*&view_type=
/facet/*?*&category=
/facet/*?*&federation_filter=
/facet/*?*&limitTo=
/facet/*?*&q=
/facet/*?*&sortBy=
/facet/*?*&tags=
/facet/*?*&view_type=

*/alt$
*/alt?
*/edit$
/*/*/*/widget_preview
/OData.svc/
/api/odata/
/analytics/add
/browse/embed
/login
/reset_password/
/tiles/
/views/INLINE/rows.json?*method=clustered2*
/api/collocate*

Tampoc poden rastrejar les següents llibreries:

/packages/
/styles/

L'user-agent "*" (o la resta d'usuaris, els que no són "gsa-crawler", "nys-crawler" i "nys-qa-crawler") no poden rastrejar els directoris anteriors però si les llibreries packages i styles.

- **Mapa del lloc web:** s'ha generat el fitxer sitemap.xml amb la funció https://www.xml-sitemaps.com/

- La seva **grandària**:

- La **tecnologia** emprada.

  En la captura de pantalla i mitjançant l'eina builtwith, que hem instal·lat, el resultat ha sigut el següent:

  Servidor web: s'utilitza Nginx, que és un servidor web/proxy invers lleuger d'alt rendiment i un proxy per a protocols de correu electrònic. Es tracta de programari lliure i de codi obert, llicenciat sota la Llicència BSD simplificada; també existeix una versió comercial distribuïda sota el nom de Nginx Plus.

  Elements de programació:

  S'utilitzen **llibreries i eines visuals o gràfiques de JavaScript** (D3, Highcharts, Javascript Infovis Toolkit). Aquestes llibreries proporcionen codi editable, inclouen gràfics en línia en el lloc web. Aquests gràfics poden ser animats, interactius i poden connectar amb les fonts de dades. Cada llibreria conté models gràfics diferents i disponibles gratuïtament (també n'inclouen de pagament).

  També s'utilitza **frameworks de JavaScript (JQuery)**, que són biblioteques que proporcionen als desenvolupadors o programadors plantilles preconstruïdes i codi JavaScript preescrit per realitzar tasques de programació estàndard. S'utilitzen per accelerar el flux de treball de desenvolupament i aplicar les millors pràctiques d'una manera fàcil i fluida.

  Llenguatge de programació utilitzat: Ruby. **Ruby** és un llenguatge de programació interpretat, reflexiu i orientat a objectes. Combina una sintaxis inspirada en Python i Perl amb característiques de programació orientada a objectes. Comparteix funcionalitat amb d'altres llenguatges de programació com Lisp, Lua, Dylan i CLU. La implementació oficial de Ruby es distribueix sota una llicència de software lliure.

  Eina web-frameworks [Ruby on Rails]: **Ruby on Rails** és un dels elements que fa que Ruby sigui tan popular. Es tracta d'un framework de Ruby que combina la simplicitat amb la possibilitat de desenvolupar aplicacions del món real amb un mínim de configuració i desenvolupant menys codi.

```
>>> builtwith.parse('https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/jj6z-iyrp/data')
{'web-servers': ['Nginx'], 'javascript-graphics': ['D3', 'Highcharts', 'Javascript Infovis Toolkit'], 'javascript-frameworks': ['jQuery'], 'web-frameworks': ['Ruby on Rails
'], 'programming-languages': ['Ruby']}
```