

Pràctica 1 – Calvo, L; Sicília, A

Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes en una web. Per a la seva realització, s'han de complir els següents punts:

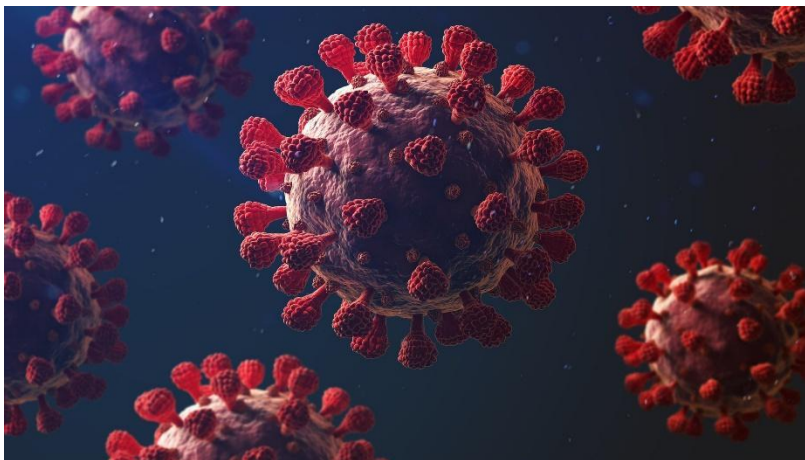
1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.
2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.
3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).
4. Representació gràfica. Presentar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.
5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.
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.
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..
8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:
 - 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
 - Other (specified above)
 - Unknown License
9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
10. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

Resolució pràctica

Exercici 1 - Context

(La Vanguardia, 2021) Tal i com és conegut, a principis de l'any 2020 es va declarar una pandèmia a escala global del SARS-COV2 a Wuhan (RP de Xina). (WHO, 2021) Com a conseqüència, a data de 26 de març del 2021 s'ha diagnosticat més de 528 mil casos i 12.596 defuncions a Catalunya. Per tal de comprendre i analitzar la situació va caldre recol·lectar un seguit de dades per tal de millorar la presa de decisions i les mesures tal i com s'observarà a posteriori a l'*Exercici 6-Agraïments*..

(BOE, 2014) Per altra banda, arran de la llei 19/2014, el govern català ha de publicar de forma pública les dades de competència pública. (GENCAT, 2021) Al declarar-se com un problema de salut pública, aquest organisme s'ha vist obligat a compartir les dades anònimament sobre l'evolució epidemiològica.



Il·lustració 1: Imatge del SARS-COV2 (Font: BBC)

(Generalitat de Catalunya, 2021) Prèviament, la Generalitat havia potenciat projectes similars de filosofia Open Data, com ara, Dades Obertes i el Fons europeu de desenvolupament regional de la Unió Europea. (Departament de Salut, 2021) En el marc d'aquest projecte es veu inclòs el projecte Dades Covid. En aquest cas, s'analitzarà

l'evolució de l'epidèmia en les principals setanta-quatre poblacions catalanes.

Exercici 2-Títol

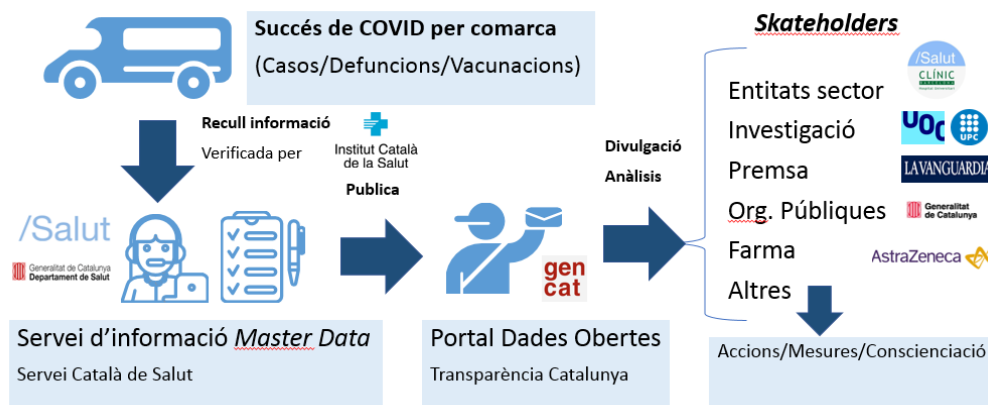
Actualització de les estadístiques sobre l'evolució del COVID a Catalunya per comarca

Exercici 3-Descripció del *dataset*

(Departament de Salut, 2021) Aquest conjunt de dades estructurades conté la informació sobre l'evolució epidemiològica amb les dades diàries de les principals ciutats catalanes. Les dades es publiquen setmanalment, per tant, l'última actualització data del 22 de març del 2021. Consta de 74 files a la data d'actualització i 13 columnes. A posteriori, es descriurà el contingut d'aquest *dataset*.

Exercici 4-Diagrames visuals

En la següent *Il·lustració 1* es pot observar l'esquema visual de l'obtenció d'informació. Un cop s'ha recollert, verificat i preprocessat les dades per població en el centre de coordinació dels centres d'atenció primària (es consideren dades com ara el número de PCR fetes, risc de rebrot, incidència acumulada o els vacunats en primera dosi), la informació és recollida pel servei d'informació de Master Data del Servei català de trànsit. (Departament de Salut, 2021) Un cop processada i verificada pels corresponents responsables, la informació es publica al portal de dades de CatSalut, Dades Covid. Aquesta informació és tractada, divulgada i/o analitzada per diferents parts interessades. Destaquen els casos d'entitats relacionades com ara l'Hospital Clínic de Barcelona, d'investigació com ara universitats, premsa, altres organitzacions interessades (com ara la Patronal de Turisme i Restauració de Catalunya o altres departaments de la Generalitat de Catalunya com la Conselleria d'Esports), empreses amb interessos (com ara farmacèutiques que estan vacunant, com Astrazeneca), entre d'altres.



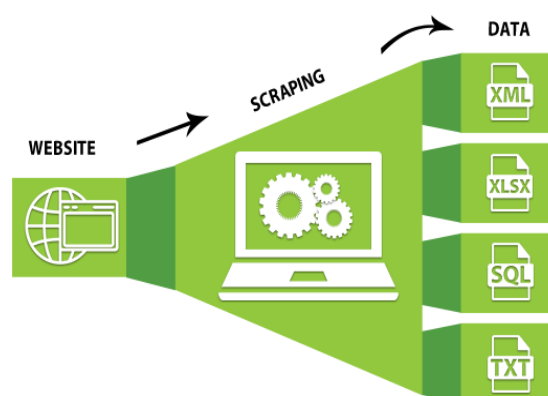
II-lustració 2: Visualització de l'obtenció d'informació

Exercici 5-Contingut

(Departament de Salut, 2021) Aquest conjunt de dades consta de 74 mostres i 13 variables, segons l'arxiu extret a 26 de març del 2021. El període de temps de les dades recollides va des del 16 de març de 2010 fins a la darrera actualització de les metadades a 22 de març del 2021. L'arxiu tal i com es pot observar en l'Annex (consultar Github annex) es pot descarregar tan en format CSV com JSON.

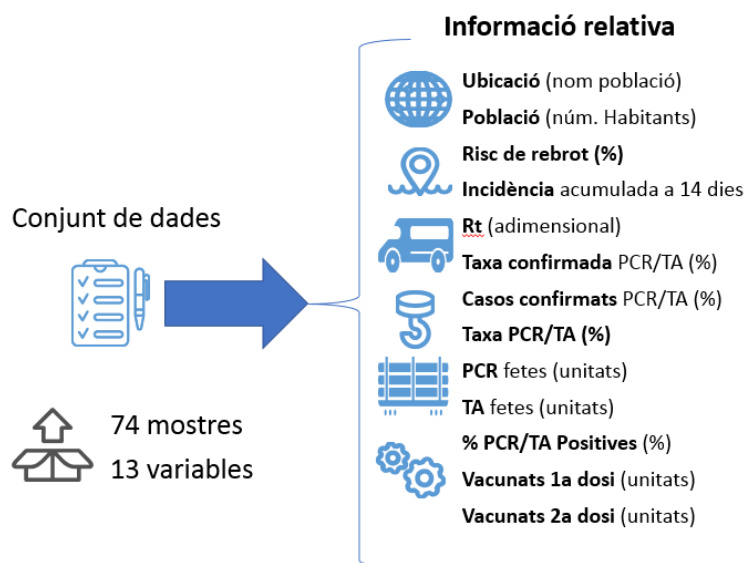
Tal i com es recull en la següent II-lustració 3, aquest conjunt de dades conté informació relativa a diferents camps. Per exemple, hi ha informació sobre el RT, número de vacunats o % PCR/TA positives.

Les dades s'han recollit mitjançant *web scraping*. A través del portal web escollit, aplicarem les tècniques d'*scraping* per tal d'obtenir les dades en el format desitjat, veure II-lustració 2. En primer lloc, s'importen les llibreries característiques com ara BeautifulSoup i observem les capçaleres per fer les peticions. A continuació, realitzem les peticions tenint en compte les bones



II-lustració 2: Esquema del web scraping

pràctiques de web scraping, com ara el *timeout* o *timesleep*. Més tard, inicialitzem log i les variables i fem la petició controlant el timeout a 10 segons. També apliquem soup, fem un log del contingut i fem la recerca de les taules. Per últim, només cal ordenar les dades tal i com ens resulta més adequat pel nostre projecte de dades i exportem el fitxer en format CSV.



Il·lustració 3: Informació sobre el conjunt de dades

En la següent *Taula 1* podem observar com és el conjunt de dades observat:

INICI	SETMANAL	DIARI	COMARQUES	MUNICIPI	MAPES	UCI	TAR	Sintomàtics	VACUNACIÓ	NOVET	COVID-19: Novel Coronavirus ... github.com	GUES	CATALÀ
Triar divisió:	Triar territori:	Triar tipus de població:											
COMARQUES	CATALUNYA	Població total											
Dades acumulades de la setmana del 16/03/2021 al 22/03/2021													
Territori	Població	Risc de rebrot	Incidència acumulada a 14 dies	Rt	Taxa confirmats per PCR/TA	Casos confirmats per PCR/TA	Taxa PCR/TA	PCR Fetes	TA Fets	% PCR/TA Positives	Vacunats 1a dosi	Vacunats 2a dosi	
AMPOSTA	21.771	20	27,56	0,85	18,37	4	1.355,01	206	89	1,48	23	326	
BADALONA	223.906	172	165,25	1,03	80,39	180	2.213,88	3.463	1.494	4,07	322	3.305	
BALAGUER	18.028	114	177,50	0,63	72,11	13	1.736,19	199	114	4,12	62	182	
BANYOLES	21.128	131	212,99	0,60	71,00	15	2.144,07	290	163	3,86	16	249	
BARBERÀ DEL VALLÈS	32.970	426	327,57	1,29	178,95	59	2.817,71	695	234	7,24	47	393	

Taula 1: (Departament de Salut, 2021) Captura d'imatge del conjunt de dades

Exercici 6-Agraïments

(Departament de Salut, 2021) Tal i com s'ha mencionat anteriorment, el conjunt de dades és de titularitat pública. En aquest cas són proveïdes pel departament d'Interior i el seu Servei Català de Salut, tot i que el propietari és la Generalitat de Catalunya.

Si fem un estat de l'art dels estudis on s'ha citat aquest conjunt de dades, (Cobarsí-Morales, 2020) destaca un informe sobre la qualitat de les dades quantitatives sobre les dades del Covid-19 al llarg del territori. Al llarg d'aquest informe es va destacar aquest informe per la qualitat i el nivell d'integritat de les dades. (Curado, Vergés, & Masferrer, 2020) No obstant, altres informes de comunicació van criticar aquesta font de dades com a causant de la confusió mèdica i d'una certa incertesa social.

Tal i com s'ha comentat en l'apartat **¡Error! No se encuentra el origen de la referencia.**, aquest conjunt de dades ha servit per tal de prendre mesures de la manera més àgil possible. (Departament de Salut, 2020) Al llarg de la publicació de la col·lecció especial COVID-19 es va citar aquest informe com a font de dades a l'hora de prendre les pertinents mesures, com ara certs confinaments, el tancament perimetral de Catalunya o d'alguns establiments.

Per altra banda, és citat en estudis científics de la malaltia, (Muñoz-Ortiz, Mompart-Penina, & Mias, 2020) com ara un anàlisi de les defuncions observades i esperades durant la pandèmia del SARS-COV2 a Catalunya. També s'ha citat aquest conjunt de dades en estudis sobre l'afectació de la Covid-19 en altres camps, com ara al consum elèctric català (Vinyets, 2020). Per últim, cal destacar que és citat en 14 publicacions i totes de l'estat espanyol.

Exercici 7-Inspiració

(Departament de Salut, 2021) La nostra intenció és poder corroborar aquesta informació, aportar més contingut i extreure'n noves conclusions de les investigacions ja realitzades. Com la majoria dels textos esmentats tenen dates de publicació antigues, ens podem plantejar diferents qüestions, com ara:

- On està augmentant l'índex de reproducció del SARS-COV2? On hi ha més risc de

rebrot? Cal prendre mesures sanitàries?

- On hi ha més casos confirmats? Hi ha relació entre el fet de tenir més població i l'índex de rebrot? Pot haver altres variables que condicionin aquesta relació?
- En quines poblacions hi ha un percentatge de PCR/Positius més elevat? S'ha d'augmentar els diagnòstics en aquelles poblacions?
- On hi ha menys vacunats en primera dosi? I en segona?
- Hi ha relació en les poblacions on hi ha més població amb el percentatge de vacunats en primera dosi? Es tracta de poblacions envellides?
- Quina població té menys percentatge de vacunats en el període de temps estudiat? Cal fer alguna campanya de suport en alguna certa geogràfica?
- El risc de rebrot coincideix amb la incidència acumulada? On coincideix? Està augmentant els casos en alguna població en concret?
- On es fan menys PCR per població? Cal fer campanyes de

Exercici 8-Llicència

(Generalitat de Catalunya, 2021) Aquest conjunt de dades té el nom de “Llicència oberta d'ús d'informació – Catalunya i és un acord de llicència que permet els usuaris compartir, modificar i utilitzar lliurement aquesta informació de manera flexible només respectant certes condicions establertes a la secció de Condicions d'ús”. (Creative Commons, 2021) Aquest tipus de llicència és del tipus *Release Under CC0: Public Domain License*.

Per tant, es permet reutilitzar la informació, distribuir i comunicar-la de manera pública i transformar la informació per a fer-ne obres derivades, per a tot el món i sense cap limitació temporal, sempre que no es contradigui amb la llicència o avís.

Exercici 9-Codi

En aquest apartat s'adjunta el codi en Python per tal de generar el conjunt de dades. El contingut del projecte també es podrà observar a

<https://github.com/aleixsf21cat/WebscrapingCovid>

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue Mar 23 21:51:42 2021

@author: Lluís
@author: Aleix
"""

from datetime import datetime
from bs4 import BeautifulSoup
import requests
import time
import re
import csv

def log_open(log_file):
    logfile = open(log_file_name, 'w')
    return logfile

def log_close(log_file):
    log_file.close()
    return

def log_print(log_file, log_str):
    # get time stamp
    dateTimeObj = datetime.now()
    # print log
    print(f'{dateTimeObj} - {log_str}', file=log_file)
    return

# set headers for requests
def set_headers():
    headers = {
        "Accept":
        "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,\
        */*;q=0.8",
        "Accept-Encoding": "gzip, deflate, sdch, br",
        "Accept-Language": "ca-ES,en-US,en;q=0.8",
        "Cache-Control": "no-cache",
        "dnt": "1",
        "Pragma": "no-cache",
        "Upgrade-Insecure-Requests": "1",
        "User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_3)
        AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87
        Safari/537.36"
    }
```



```
    return headers

# do_requests
def do_request(url, headers, timesleep, timeout, retries):
    log_print(log, "do_request - Downloading: " + url)
    download = False
    count = retries
    while (download == False):
        try:
            page = requests.get(url, headers=headers, timeout=10)
        except requests.exceptions.Timeout:
            log_print(log, "Requests - requests.exceptions.Timeout")
            time.sleep(timesleep)
        except requests.exceptions.RequestException:
            log_print(log, "Requests - requests.exceptions.RequestException")
            time.sleep(timesleep)
        if page != None:
            download = True
            count -= 1
            if count == 0:
                break
    return page

if __name__ == "__main__":

    # Inicialització log
    log_file_name = "scraper.log"
    log = log_open(log_file_name)
    log_print(log, "Inici")

    # Inicialització variables
    log_print(log, "Inicialitzant variables globals - Inici")
    url = "https://dadescovid.cat/municipi?tipus_territori=territori"
    log_print(log, "url: " + url)
    # requests parameters
    timeout = 30
    log_print(log, "timeout: " + str(timeout))
    timesleep = 3
    log_print(log, "timesleep: " + str(timesleep))
    retries = 5
    log_print(log, "retries: " + str(retries))
    log_print(log, "Inicialitzant variables globals - Fi")

    # set headers
    headers = set_headers()

    # request
    log_print(log, "requests - Inici")

    # Requests controlant el timeout = 10 segons
```

```
page = do_request(url, headers, timesleep, timeout, retries)
log_print(log, "requests - Fi")
log_print(log, "Status Code: " + str(page.status_code))

# soup
soup = BeautifulSoup(page.content, "html.parser")
log_print(log, "soup.name: " + str(soup.name))

# log del contingut
log_print(log, "sop.prettify()")
log_print(log, soup.prettify())

# Recerca de taules.
log_print(log, "Recerca de tautes")
for table in soup.find_all('table'):
    log_print(log, table.get_text())

#for tr in soup.find_all('tr')[2:]:
#    tds = tr.find_all('td')
#    print(tds[0].get_text())
#    print(tds[1].get_text())

table = soup.find("table", attrs={"class": "table center"})
#print(table)
headings = []
#for th in table.find_all("th"):
#    for td in th.find_all("td"):
#        headings.append(td.text.replace('\n', ' ').strip())
#print(headings)
log_print(log, "headings")
headings = []
for ths in table.find_all("th"):
    log_print(log, "ths.get_text(): " + ths.text.replace('\n', ' '))
    headings.append(ths.text.replace('\n', ' ').strip())

log_print(log, "headings-summary: ")
log_print(log, headings)

table_data = []
for trs in table.tbody.find_all("tr"):
    t_row = []
    for tds in trs.find_all("td"):
        log_print(log, "tds")
        element = re.sub("(\xa0)|(\n)|,", "", tds.text)
        t_row.append(element)
    table_data.append(t_row)

log_print(log, "table_data: ")
log_print(log, table_data)
```

```
log_print(log, "csv_output: ")
outfile = open("table_data.csv", "w", newline='')
writer = csv.writer(outfile)
writer.writerow(headings)
for row in table_data:
    log_print(log, row)
    writer.writerow(row)

# Finalització log
log_print(log, "Final")
log_close(log)
```


Exercici 10-Generació del dataset

En aquest apartat s'adjunta una captura d'imatge de l'arxiu resultant en format CSV. Tal i com s'havia mostrat prèviament es mostren diferents camps com ara el territori, població, risc de rebrot, incidència acumulada a 14 dies, RT, Taxa de confirmats per PCR/TA, Casos confirmats per PCR/TA, % PCR/TA Positives, Vacunats 1a dosi i Vacunats 2a dosi.

Territori,Població,A,Risc de rebrot,Incidència acumulada a 14 dies,Rt,Taxa confirmats per PCR/TA,Casos confirmats per PCR/TA,Taxa PCR/TA,PCR Fetes,TA Fets,% PCR/TA Positives,Vacunats 1a dosi,Vacunats 2a dosi
AMPOSTA,21.771,26,3215,092,1837,4,1.11157,164,78,183,118,257
BADALONA,223.906,196,17016,114,9290,208,2.09374,3.288,1.400,494,906,2.607
BALAGUER,18.028,116,18305,062,5547,10,1.50322,164,107,496,83,205
BANYOLES,21.128,254,26032,095,12779,27,3.12382,497,163,420,66,214
BARBERÀ DEL VALLÀS,32.970,475,35487,132,20625,68,3.24537,842,228,697,185,383
BARCELONA,1.647.581,239,20126,118,10852,1.788,2.2314,25.986,10.642,540,10.052,23.283
BLANES,39.749,339,21133,159,13585,54,1.91200,555,205,826,174,388
CALAFELL,25.743,50,3496,142,1942,5,1.62374,330,88,128,107,60
CAMBRIJLS,34.443,97,7549,127,3484,12,94939,222,105,458,176,324
CANET DE MAR,14.565,191,13045,141,4119,6,2.60899,291,89,189,38,89
CANOVELLES,17.181,321,40743,079,12805,22,2.86363,339,153,470,57,141
CASTELLAR DEL VALLÀS,24.196,189,17358,108,8679,21,3.04182,550,186,339,116,255

Il·lustració 4: Captura d'imatge del fitxer resultant en format .CSV

La publicació DOI té el següent registre: [10.5281/zenodo.464268](https://doi.org/10.5281/zenodo.464268)



tfg.aleix@gmail.com

March 28, 2021
Dataset
Open Access
Edit
New version
0 views
0 downloads
See more details...
Indexed in
OpenAIRE
Publication date: March 28, 2021
DOI: 10.5281/zenodo.464268
Keyword(s): COVID
License (for files): Creative Commons Attribution 4.0 International

Evolució Epidemiològica de les dades del COVID a Catalunya de forma setmanal

Aleix Sicilia Fuentes; Lluís Calvo

(Departament de Salut, 2021) Aquest conjunt de dades estructurades conté la informació sobre l'evolució epidemiològica amb les dades diàries de les principals ciutats catalanes. Les dades es publiquen setmanalment, per tant, l'última actualització data del 22 de març del 2021. Consta de 74 files a la data d'actualització i 13 columnes. A posteriori, es descriurà el contingut d'aquest dataset. (Departament de Salut, 2021) Aquest conjunt de dades estructurades conté la informació sobre l'evolució epidemiològica amb les dades diàries de les principals ciutats catalanes. Les dades es publiquen setmanalment, per tant, l'última actualització data del 22 de març del 2021. Consta de 74 files a la data d'actualització i 13 columnes. A posteriori, es descriurà el contingut d'aquest dataset.

Referències

Departament de Salut. (2021). *Dades COVID*. Barcelona.

Incidència acumulada a 14 dies	Rt	Taxa confirmats per PCR/TA	Casos confirmats per PCR/TA	Taxa PCR/TA	PCR Fets	TA Fets	% PCR/TA Positives	Vacunats 1a dosi	Vacunats 2a dosi
3215	092	1837	4	1.11157	164	78	183	118	257
17016	114	9290	208	2.09374	3.288	1.400	494	906	2.607
18305	062	5547	10	1.50322	164	107	496	83	205

Il·lustració 5: Captura d'imatge de la publicació a Zenodo amb el registre 10.5281/zenodo.464268

Referències

- BOE. (2014). *Ley 19/2014, de 29 de diciembre, de transparencia, acceso a la información y buen gobierno*. Madrid: Legislación consolidada del gobierno de España.
- Cobarsí-Morales, J. (2020). DOI: <https://doi.org/10.3145/thinkepi.2020.e14d02>: UOC.
- Creative Commons. (01 / Febrer / 2021). *CC0 1.0 Universal*. Recollit de <https://creativecommons.org/publicdomain/zero/1.0/legalcode>
- Curado, B., Vergés, G., & Masferrer. (2020). *Confusió mèdica i incertesa social: apunts etnogràfics sobre la construcció del risc durant la pandèmia per la Covid-19 a Catalunya*. Bellaterra: Revistes UAB.
- Departament de Salut. (2020). *Noves mesures contra la COVID-19 a Catalunya*. Barcelona: Material de divulgació. Col·lecció especial COVID-19.
- Departament de Salut. (2021). *Dades COVID*. Barcelona.
- GENCAT. (26 / març / 2021). *Últimes dades Coronavirus*. Recollit de <https://aguas.gencat.cat/ca/actualitat/ultimes-dades-coronavirus/mapa-per-municipis/>
- Generalitat de Catalunya. (Març / 2021). Recollit de Dades Obertes: http://governobert.gencat.cat/ca/dades_obertes/
- Generalitat de Catalunya. (18 / Març / 2021). *Accidents de trànsit amb morts o ferits greus a Catalunya*. Recollit de Portal de transparència: <https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb>
- La Vanguardia. (26 / 03 / 2021). Restricciones por el COVID. *La Vanguardia*, p. 2021. Recollit de <https://www.lavanguardia.com/local/barcelona/20210326/6607928/restricciones-catalunya-barcelona-girona-lleida-tarragona-ultimas-noticias-26-marzo-hoy-en->

directo.html

Muñoz-Ortiz, L., Mompert-Penina, A., & Mias, M. (2020). *Anàlisi de les defuncions observades i esperades durant la pandèmia de la Covid-19 a Catalunya*. 84.88.27.52: UPC Commons.

Vinyets, J. (2020). *Estudi de l'afectació de la Covid-19 al consum elèctric català*. Barcelona: UPC Commons.

WHO. (2021). *Data SARS-COV2*. New York.

Contribució al projecte

Al llarg d'aquest projecte, tots els apartats s'han realitzat de manera conjunta entre els integrants d'aquest grup de treball

Contribucions	Signa
Recerca prèvia	Calvo, L; Sicília, A
Redacció de les respostes	Calvo, L; Sicília, A
Desenvolupament codi	Calvo, L; Sicília, A
Generació Github	Calvo, L; Sicília, A
Generació DOI	Calvo, L; Sicília, A