Marc Alemany (malemanys)
Josep Garcia (josepgarcia)

Apartat 1 [0,25]

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

Degut al gran impacte que ha tingut en les nostres vides el COVID-19 hem decidit buscar les dades de casos, proves, vacunes, persones ingressades i mortes a Catalunya.

Hem triat la web: https://dadescovid.cat/diari perquè és una font d'informació oficial que actualitza diàriament les dades i ens dona la seguretat de poder-les extreure permanentment sense haver de dependre d'una font privada que pugui tallar-nos el servei.

Apartat 2 [0,25]

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

El títol que hem triat pel dataset és Evolució del COVID a Catalunya, ja que és exactament el que estudiarem en el dataset.

Apartat 3 [0,25]

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).

Tal com expressa el títol, el dataset està basat en les dades més importants a tenir en compte per poder observar i avaluar la evolució del COVID a Catalunya des de l'1 de Març del 2020 fins a l'actualitat.

Tenim dades sobre els casos confirmats, nombre de proves fetes, nivell de vacunació (faltaria incloure la població catalana per saber el percentatge en què ens trobem per exemple), el nivell de saturació de la sanitat catalana gràcies als Ingressats o bé les conseqüències a nivell de Defuncions. Per tant, podem utilitzar el dataset per avaluar la situació des de molts punts de vista.

Les dades obtingudes són diàries on el primer cap seria el dia que fa referència les dades i la resta són valors numèrics indicant la magnitud de la qual fa referència la variable.

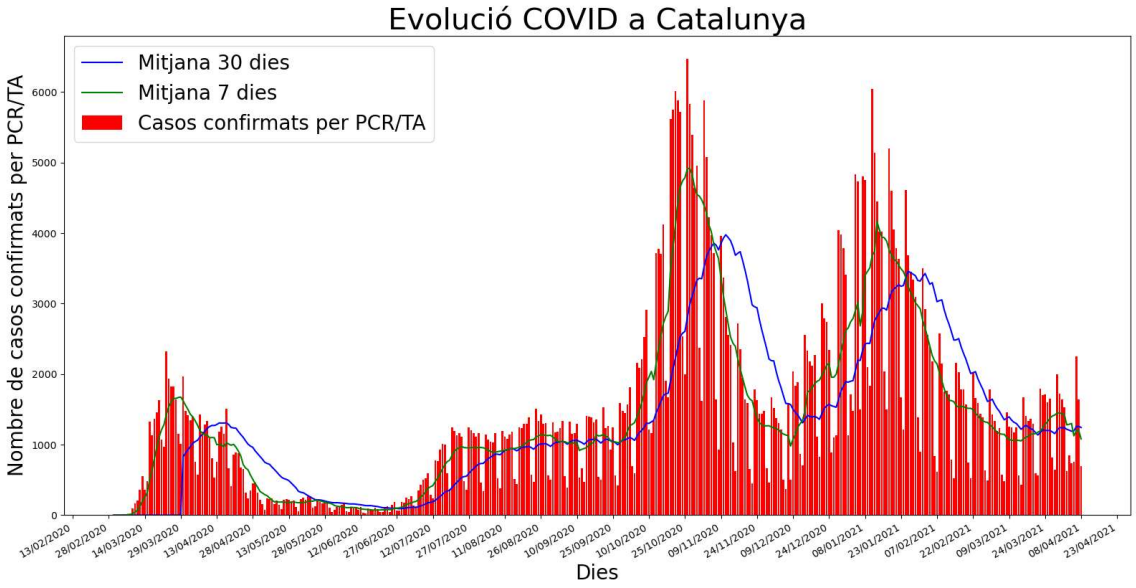
Seran necessàries varies actuacions per poder treballar el dataset correctament, les explicarem a l'apartat 5.

El resultat final del dataset serà un CSV per poder-lo tractar i/o visualitzar com es necessiti.

Marc Alemany (malemanys)
Josep Garcia (josepgarcia)

Apartat 4 [0,25]

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



Apartat 5 [1]

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

El conjunt de dades ens mostra diàriament 9 valors relacionats amb el COVID:

1. Data
2. Casos confirmats per PCR/TA
3. PCR Fetes
4. TA Fets
5. % PCR/TA Positives
6. Vacunats 1a dosi
7. Vacunats 2a dosi
8. Ingressats
9. Defuncions

Les dades comencen a recollir-se el dia 01/03/2020 amb l'inici de la pandèmia fins a l'actualitat i la seva actualització és diària.

S'han recollit utilitzant web scraping, mitjançant les llibreries BeautifulSoup i requests.

La extracció està dividida en dues parts:

1. Títols
2. Dades

Els títols els hem extret buscant l'etiqueta HTML 'th' i hem extret el text mitjançant un bucle que afegeix les dades a la variable 'titles'.

Les dades les hem extret buscant les etiquetes 'tr' i 'td' i les hem extret mitjançant un doble bucle per generar una llista per cada fila i els elements de cada fila.

Finalment, hem inclòs les dades a un dataframe on inicialment hem inclòs els títols i posteriorment les dades.

Mostra del dataset inicial:

Out[2]:

	Data	Casos confirmats per PCR/TA	PCR Fetes	TA Fets	% PCR/TA Positives	Vacunats 1a dosi	Vacunats 2a dosi	Ingressats	Defuncions
0	08/04/2021*	698	4.063	4.645	6,96	59.361	3.061	1.695	0
1	07/04/2021*	1.644	20.691	7.593	7,15	80.726	4.369	1.721	19
2	06/04/2021*	2.255	20.780	10.205	8,34	46.243	2.264	1.726	21
3	05/04/2021	761	5.210	3.165	9,31	17.414	7	1.721	16
4	04/04/2021	740	5.418	2.827	10,01	12.592	3	1.617	13
...	...	...	...	...	...	...	...	...	...
399	05/03/2020	18	37	1	13,89	0	0	0	0
400	04/03/2020	7	29	0	3,45	0	0	0	0
401	03/03/2020	9	25	0	0	0	0	0	0
402	02/03/2020	5	35	1	8,33	0	0	0	0
403	01/03/2020	2	22	0	0	0	0	0	0

404 rows × 9 columns

Una vegada hem tingut les dades, hem visc convenient tractar-les per poder realitzar uns càlculs que s'inclouran en un dataframe final, el tractament ha estat:

1. Ordenar les dades de forma ascendent on primer veiem els valors amb la data més antiga i els últims valors tenen la data més actual.
2. Hem tret l'asterisc "*" als valors que ho continguin del camp "data" per tenir un camp homogeni. Aquest asterisc indica que les dades no són definitives però a l'actualitzar-se diàriament hem vist un avantatge treure-ho per si volguéssim operar amb les dates.
3. Convertir el camp 'Data' a format datetime per poder operar amb les dates si fos necessari, en aquest cas ho hem aplicat per mostrar correctament l'eix de les X del gràfic.
4. Convertim la resta de valor a enters (eliminant el punt) excepte el camp % PCR/TA Positives que serà un real (substituïm coma per punt).

Una vegada tractades les dades, hem realitzat varis càlculs que ens serviran per poder veure la evolució de les dades amb mitjanes mòbils, variacions diàries o de 7D o el % de les variacions.

Com molts dels càlculs obtinguts no es poden calcular els primers dies i el resultat obtingut és N/A hem optat per substituir-los per 0, i així podrem homogeneïtzar el 'type'.

En els casos necessaris hem arrodonit a 2 decimals perquè la seva visualització sigui òptima.

Realitzades les tasques anteriors, hem obtingut un dataframe amb els camps següents:

1. Data
2. Casos confirmats per PCR/TA
3. PCR Fetes
4. TA Fets
5. % PCR/TA Positives
6. Vacunats 1a dosi
7. Vacunats 2a dosi
8. Ingressats
9. Defuncions
10. Canvi_diari
11. Canvi_7D
12. Canvi_%
13. Canvi_7D_%
14. MA_7
15. MA_30

Mostra del dataset final:

Out[3]:

	Data	Casos confirmats per PCR/TA	PCR Fetes	TA Fets	% PCR/TA Positives	Vacunats 1a dosi	Vacunats 2a dosi	Ingressats	Defuncions	Canvi_diari	Canvi_7D	Canvi_%	Canvi_7D_%	MA_7	MA_30
0	2020-03-01	2	22	0	0.00	0	0	0	0	0	0	0.00	0.00	0.00	0.00
1	2020-03-02	5	35	1	8.33	0	0	0	0	3	0	1.50	0.00	0.00	0.00
2	2020-03-03	9	25	0	0.00	0	0	0	0	4	0	0.80	0.00	0.00	0.00
3	2020-03-04	7	29	0	3.45	0	0	0	0	-2	0	-0.22	0.00	0.00	0.00
4	2020-03-05	18	37	1	13.89	0	0	0	0	11	0	1.57	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
399	2021-04-04	740	5418	2827	10.01	12592	3	1617	13	-111	95	-0.13	0.15	1301.29	1189.97
400	2021-04-05	761	5210	3165	9.31	17414	7	1721	16	21	-1240	0.03	-0.62	1124.14	1196.13
401	2021-04-06	2255	20780	10205	8.34	46243	2264	1726	21	1494	529	1.96	0.31	1199.71	1255.30
402	2021-04-07	1644	20691	7593	7.15	80726	4369	1721	19	-611	6	-0.27	0.00	1200.57	1261.43
403	2021-04-08	698	4063	4645	6.96	59361	3061	1695	0	-946	-834	-0.58	-0.54	1081.43	1242.97

404 rows × 15 columns

Apartat 6 [1,5]

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.

El propietari de les dades és la Generalitat de Catalunya que permet la reutilització dels continguts i de les dades sempre que se'n citi la font i la data d'actualització i que no es

desnaturalitzi la informació (article 8 de la Llei 37/2007) i també que no es contradigui amb una llicència específica.

La mateixa web presenta un estudi amb una gran segmentació de dades per regions o comarques, amb càlculs setmanals de l'impacte del COVID amb indicadors com el Risc de rebrot, Rt o Taxa de confirmats per exemple.

El nostre objectiu ha estat ser capaços de fer el webscraping correctament i realitzar càlculs que no han fet com són les mitjanes mòbils de X dies que indiquen una tendència de les dades o els percentatges de variació de X dies també.

En quan a la informació de l'arxiu robots.txt no l'hem pogut valorar perquè no tenen aquest arxiu:

Per tant, hem entès que amb l'avís legal de la part inferior de la web ens permetien reutilitzar els continguts:

Apartat 7 [1,5]

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.

Aquest conjunt de dades és interessant perquè pot donar resposta a varies preguntes:

1. Quina és la evolució de la pandèmia a Catalunya ?
2. Es troba Catalunya en una zona de perill creixent o decreixent ?
3. Quina és la evolució de la pressió hospitalària ?
4. Quina és la evolució de la vacunació ?
5. Quin percentatge de casos acaben morint ?

Ens podríem fer moltíssimes preguntes amb aquest conjunt de dades on enfocant els càlculs cap a un àmbit o un altre podríem treure moltes conclusions.

Nosaltres ens hem centrat en la evolució dels casos on veiem clarament els diferents pics de la corba de contagis que hem viscut.

Previsiblement estarem davant d'un altre increment de casos, ja que en anteriors situacions on el Govern ha afluixat les mesures anti-covid per poder gaudir de vacances, i per tant, incrementar el contacte social, s'ha vist incrementada la corba de contagis directament proporcional (no estem valorant si ha estat una decisió encertada o no, ens limitem a descriure les dades).

Per tant, estem davant d'un conjunt de dades molt interessant on la seva evolució anirà lligada a la nostra vida actual i dels pròxims mesos, esperem que no siguin anys.

Apartat 8 [1]

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

La llicència seleccionada és la següent:

○ Released Under CC BY-NC-SA 4.0 License

Hem triat aquesta llicència perquè creiem que hem fet una bona feina i el resultat pot ser útil per alguna persona que vulgui aprofundir en les dades extretes, fet que la llicència permet compartir i adaptar el material.

Tot i així, com la font de les dades no som nosaltres, sinó de la Generalitat de Catalunya, ens volem assegurar de 3 punts que veiem importants:

1. Reconeixement: S'ha de reconèixer l'autoria de la forma més apropiada, proporcionar un enllaç a la llicència i indicar si s'ha realitzat algun canvi. Així ens assegurem que indiquin la Generalitat de Catalunya.
2. No Comercial: No es pot utilitzar el material per fins comercials.
3. Compartir Igual: Si hi ha canvis i es volen difondre els resultats, s'ha de fer amb la mateixa llicència, així ens assegurem que no serà comercial i sempre existirà el reconeixement de l'autoria.

Apartat 9 [2]

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

Està publicat al GitHub.

Marc Alemany (malemanys)
Josep Garcia (josepgarcia)

Apartat 10 [2]

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

Està publicat al GitHub.

Participació dels integrants en cada apartat:

Contribucions	Signa
Recerca prèvia	Josep Garcia Gutiérrez, Marc Alemany Selle
Redacció de les respostes	Josep Garcia Gutiérrez, Marc Alemany Selle
Desenvolupament codi	Josep Garcia Gutiérrez, Marc Alemany Selle