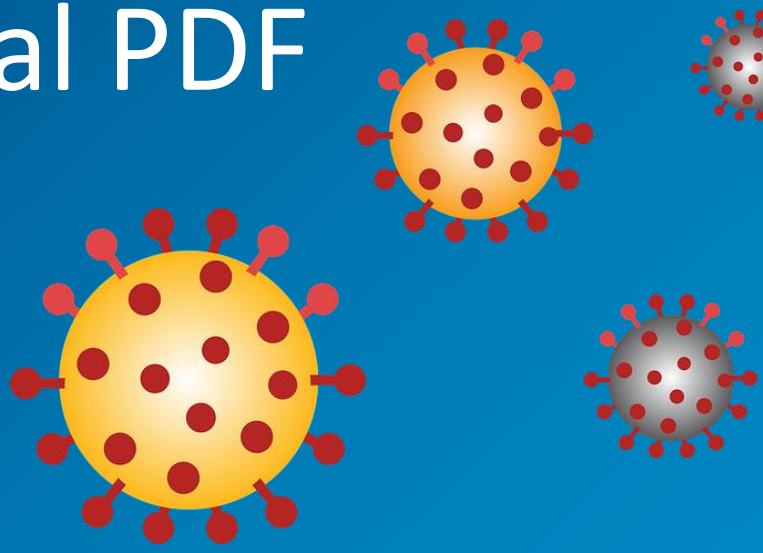
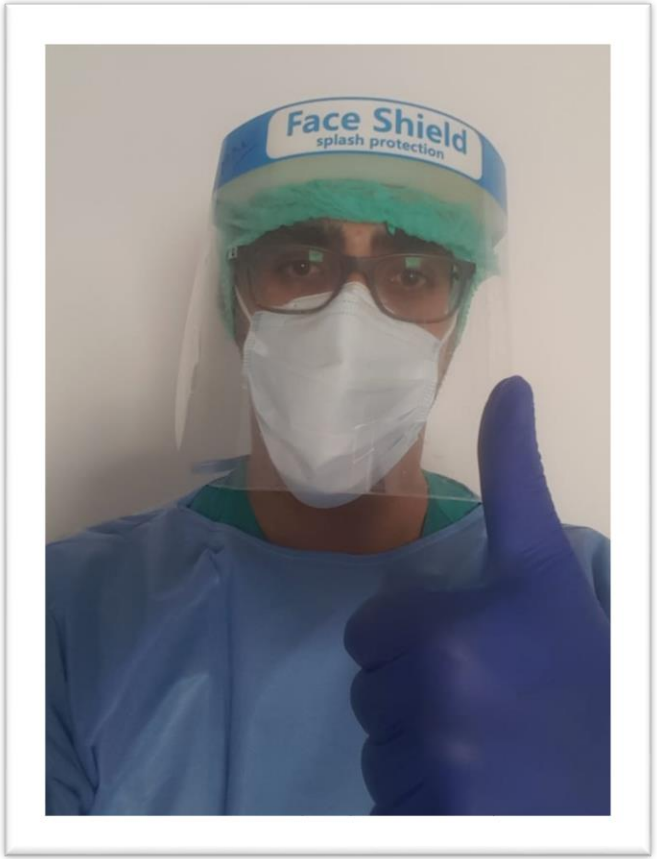
How to get COVID19 data from official PDF with Python



Víctor Vicente Palacios
Clinical Data Scientist
Philips Healthcare

Lockdown, and now what can I do?



COVID19

@victorvicpal

I am a Data Scientist so…

COVID19

@victorvicpal

Let’s find some data

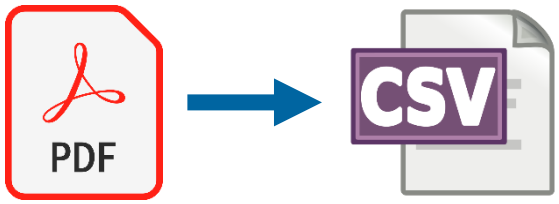
2. SITUACIÓN ACTUAL

Situación en España:

En España, hasta el momento se han registrado 237 casos y 3 fallecidos. (Tabla 1 y Figura 1).

Tabla 1. Casos COVID-19, incidencia acumulada (IA), ingreso en UCI y fallecidos por Comunidades autónomas en España, 05.03.2020.

CCAA	Total casos	IA (casos/100.000 habitantes)	Ingreso en UCI	Fallecidos
Andalucía	12	0,14	1	0
Aragón	1	0,08	1	0
Asturias	5	0,49	2	0
Baleares	6	0,52	0	0
Canarias	8	0,37	0	0
Cantabria	10	1,72	0	0
Castilla La Mancha	13	0,64	1	0
Castilla y León	11	0,46	1	0
Cataluña	24	0,31	0	0
Ceuta	0	0,00	0	0
C. Valenciana	19	0,38	0	1
Extremadura	6	0,56	0	0
Galicia	1	0,04	0	0
Madrid	90	1,35	2	1
Melilla	0	0,00	0	0
Murcia	0	0,00	0	0
Navarra	3	0,46	1	0
País Vasco	17	0,77	0	1
La Rioja	11	3,47	0	0
Total	**237**	**0,51**	**9**	**3**



tabula-py

build passing | pypi package 2.1.0 | docs passing | patreon donate

build passing | coverage 48%

tika-python

A Python port of the Apache Tika library

COVID19

@victorvicpal

Data…?

```
from tika import parser

raw = parser.from_file(inputfile)

raw['content']
```

“Centro de Coordinación de Alertas y Emergencias Sanitarias \n 1 SECRETARIA GENERAL DE SANIDAD DIRECCIÓN GENERAL DE SALUD PÚBLICA, CALIDAD E INNOVACIÓN \n Actualización nº 87. \n Enfermedad por el coronavirus (COVID-19). 26.04.2020 (datos consolidados a las 21:00 horas del 25.04.2020) SITUACIÓN EN ESPAÑA En España, hasta el momento se han notificado un total de 207.634 casos confirmados de COVID.19 por PCR, 23.190 fallecidos y 98.732 curados (Tabla 1, Tabla 2, Figura 1 y Figura 2)…”

String > Data

```
def get_ccaa_tables(string, keywords):
    tabs = []
    for kw in keywords:
        i1 = string.find(kw)
        i1 = i1 + string[i1:].find('Andalucía')
        i2 = i1 + string[i1:].find('ESPAÑA')
        tabs.append(string[i1:i2])
    return tabs
```

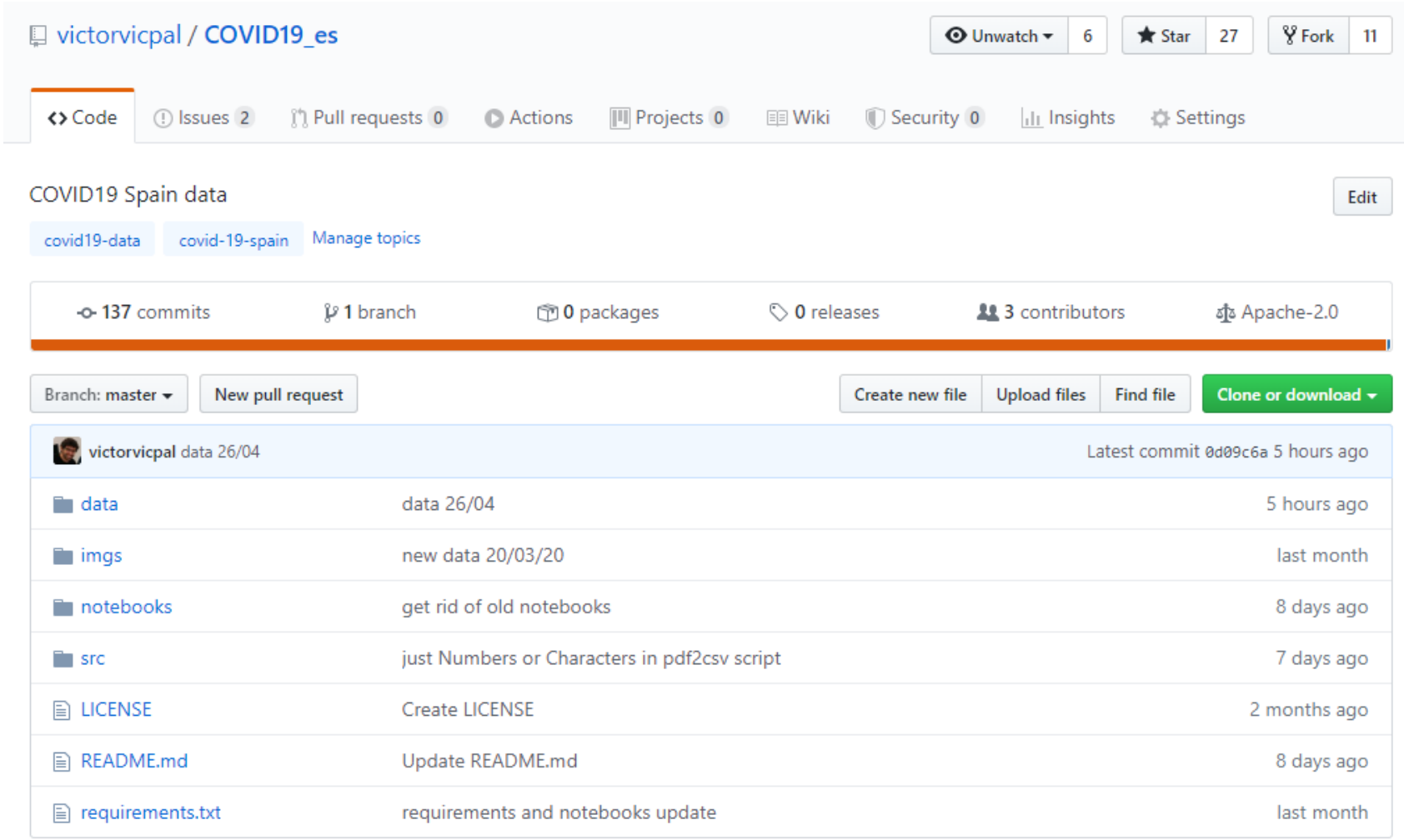
```
[['Andalucía', '11774', '71', '0.6', '21.01', ''],
 ['Aragón', '4955', '33', '0.7', '67.08', ''],
 ['Asturias', '2249', '11', '0.5', '34.90', ''],
 ['Baleares', '1854', '7', '0.4', '27.84', ''],
 ['Canarias', '2167', '12', '0.6', '11.56', ''],
 ['Cantabria', '2083', '12', '0.6', '56.96', ''],
 ['CastillaLaMancha', '15609', '100', '0.6', '94.01', ''],
 ['CastillayLeón', '16222', '232', '1.5', '171.03', ''],
 ['Cataluña', '46811', '550', '1.2', '166.56', ''],
 ['Ceuta', '100', '0', '0.0', '8.26', ''],
 ['CValenciana', '10160', '94', '0.9', '26.36', ''],
 ['Extremadura', '2736', '18', '0.7', '14.61', ''],
 ['Galicia', '9176', '60', '0.7', '68.16', ''],
 ['Madrid', '59126', '307', '0.5', '188.18', ''],
 ['Melilla', '110', '2', '1.9', '13.87', ''],
 ['Murcia', '1474', '6', '0.4', '1.67', ''],
 ['Navarra', '4712', '85', '1.8', '113.57', ''],
 ['PaísVasco', '12455', '89', '0.7', '76.23', ''],
 ['LaRioja', '3861', '40', '1.0', '183.71', '']]
```

String > Data

```
1 data[cols]
```

	CCAA	fecha	casos	nuevos	incr %	IA	Hospitalizados	HospitalizadosNuevos	UCI	UCINuevos	muertes	muertesNuevos
0	Andalucía	26.04.2020	11774	71	0.6	21.01	5748	33	716	0	1145	14
1	Aragón	26.04.2020	4955	33	0.7	67.08	2395	23	257	2	712	3
2	Asturias	26.04.2020	2249	11	0.5	34.90	1790	30	134	2	249	10
3	Baleares	26.04.2020	1854	7	0.4	27.84	1062	4	166	0	175	1
4	Canarias	26.04.2020	2167	12	0.6	11.56	880	2	171	0	131	1
5	Cantabria	26.04.2020	2083	12	0.6	56.96	991	9	78	0	183	1
6	CastillaLaMancha	26.04.2020	15609	100	0.6	94.01	8417	32	565	6	2330	38
7	CastillayLeón	26.04.2020	16222	232	1.5	171.03	7653	98	515	4	1666	27
8	Cataluña	26.04.2020	46811	550	1.2	166.56	25665	1245	2583	7	4566	68
9	Ceuta	26.04.2020	100	0	0.0	8.26	10	0	4	0	4	0
10	CValenciana	26.04.2020	10160	94	0.9	26.36	5013	35	654	1	1186	14
11	Extremadura	26.04.2020	2736	18	0.7	14.61	1480	33	108	1	422	6
12	Galicia	26.04.2020	9176	60	0.7	68.16	2735	13	85		394	6
13	Madrid	26.04.2020	59126	307	0.5	188.18	5892		873		7922	74
14	Melilla	26.04.2020	110	2	1.9	13.87	44	0	3	0	2	0
15	Murcia	26.04.2020	1474	6	0.4	1.67	627	2	106	1	127	1
16	Navarra	26.04.2020	4712	85	1.8	113.57	1942	5	129	0	431	2
17	PaísVasco	26.04.2020	12455	89	0.7	76.23	6426	51	533	8	1230	18
18	LaRioja	26.04.2020	3861	40	1.0	183.71	1380	20	84	1	315	4

Data > Github



victorvicpal / COVID19_es

Unwatch 6 | Star 27 | Fork 11

Code | Issues 2 | Pull requests 0 | Actions | Projects 0 | Wiki | Security 0 | Insights | Settings

COVID19 Spain data

Edit

covid19-data | covid-19-spain | Manage topics

137 commits | 1 branch | 0 packages | 0 releases | 3 contributors | Apache-2.0

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

victorvicpal data 26/04 — Latest commit 0d09c6a 5 hours ago

data	data 26/04	5 hours ago
imgs	new data 20/03/20	last month
notebooks	get rid of old notebooks	8 days ago
src	just Numbers or Characters in pdf2csv script	7 days ago
LICENSE	Create LICENSE	2 months ago
README.md	Update README.md	8 days ago
requirements.txt	requirements and notebooks update	last month

Format changes!!

2. SITUACIÓN ACTUAL

Situación en España:

En España, hasta el momento se han registrado 237 casos y 3 fallecidos. (Tabla 1 y Figura 1).

Tabla 1. Casos COVID-19, incidencia acumulada (IA), ingreso en UCI y fallecidos por Comunidades autónomas en España, 05.03.2020.

CCAA	Total casos	IA (casos/100.000 habitantes)	Ingreso en UCI	Fallecidos
Andalucía	12	0,14	1	0
Aragón	1	0,08	1	0
Asturias	5	0,49	2	0
Baleares	6	0,52	0	0
Canarias	8	0,37	0	0
Cantabria	10	1,72	0	0
Castilla La Mancha	13	0,64	1	0
Castilla y León	11	0,46	1	0
Cataluña	24	0,31	0	0
Ceuta	0	0,00	0	0
C. Valenciana	19	0,38	0	1
Extremadura	6	0,56	0	0
Galicia	1	0,04	0	0
Madrid	90	1,35	2	1
Melilla	0	0,00	0	0
Murcia	0	0,00	0	0
Navarra	3	0,46	1	0
País Vasco	17	0,77	0	1
La Rioja	11	3,47	0	0
Total	**237**	**0,51**	**9**	**3**

COVID19

@victorvicpal

Format changes!!

Fernando Simon got sick

CCAA	TOTAL confirmados*	IA (14 d.)	Casos que han precisado hospitalización*	Casos que han precisado ingreso en UCI	Fallecidos	Curados	*Nuevos*
Andalucía	**5.818**	61,03	2.867	235	248	160	*413*
Aragón	**2.272**	156,52	1.176	165	138	204	*194*
Asturias	**1.236**	101,97	529	65	55	90	*78*
Baleares	**1.069**	85,00	399	85	42	111	*69*
Canarias	**1.262**	51,73	483	94	55	57	*58*
Cantabria	**1.171**	191,54	522	50	37	35	*71*
Castilla La Mancha	**6.424**	288,12	3.225	344	708	296	*566*
Castilla y León	**6.211**	240,88	2.601	325	516	1.028	*410*
Cataluña	**18.773**	226,43	12.974	1.652	1.672	4.966	*2.616*
Ceuta	**34**	38,93	3	3	1	0	*9*
C. Valenciana	**5.508**	99,27	2.189	356	339	200	*398*
Extremadura	**1.628**	138,15	371	51	133	91	*68*
Galicia	**4.039**	138,80	1.250	149	84	187	*316*
Madrid	**27.509**	339,74	15.140	1.514	3.603	9.330	*3.419*
Melilla	**54**	42,78	27	3	1	0	*3*
Murcia	**974**	58,71	283	59	34	20	*35*
Navarra	**2.305**	304,49	1.035	99	113	192	*159*
País Vasco	**6.320**	251,61	3.594	307	325	1.796	*263*
La Rioja	**1.810**	459,28	575	51	85	496	*77*
ESPAÑA	**94.417**	177,01	**49.243**	**5.607**	**8.189**	**19.259**	*9.222*

IA: Incidencia acumulada (casos acumulados por 100.000 habitantes)

COVID19

@victorvicpal

Format changes!!

Fernando Simon came back

Different tables

More info

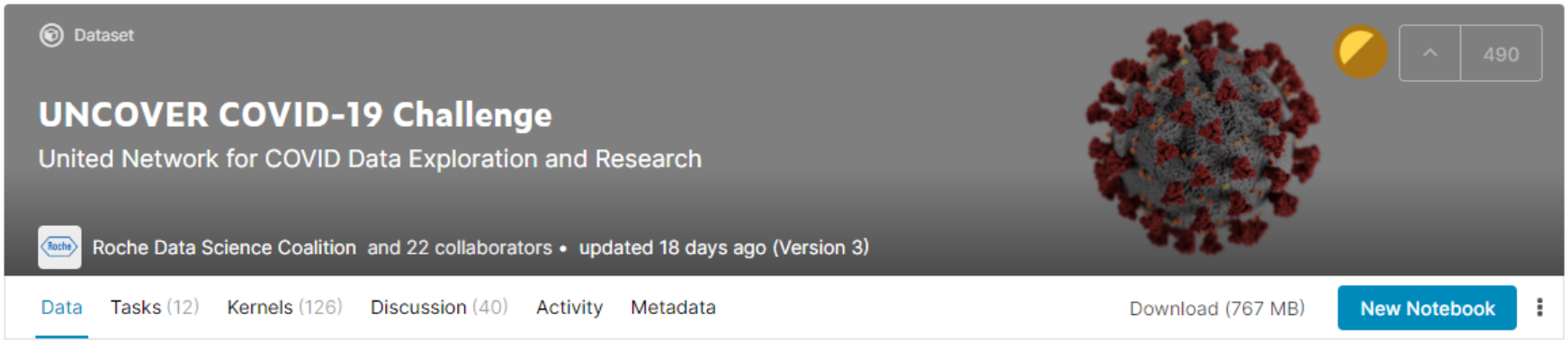
Confirmados por PCR				
CCAA	Total	*Nuevos*	Incremento confirmados	IA (14 d.)
Andalucía	11.774	*71*	0,6%	*21,01*
Aragón	4.955	*33*	0,7%	*67,08*
Asturias	2.249	*11*	0,5%	*34,90*
Baleares	1.854	*7*	0,4%	*27,84*
Canarias	2.167	*12*	0,6%	*11,56*
Cantabria	2.083	*12*	0,6%	*56,96*
Castilla La Mancha	15.609	*100*	0,6%	*94,01*
Castilla y León	16.222	*232*	1,5%	*171,03*
Cataluña	46.811	*550*	1,2%	*166,56*
Ceuta	100	*0*	0,0%	*8,26*
C. Valenciana	10.160	*94*	0,9%	*26,36*
Extremadura	2.736	*18*	0,7%	*14,61*
Galicia*	*9.176*	*60*	0,7%	*68,16*
Madrid	59.126	*307*	0,5%	*188,18*
Melilla	110	*2*	1,9%	*13,87*
Murcia	1.474	*6*	0,4%	*1,67*
Navarra	4.712	*85*	1,8%	*113,57*
País Vasco	12.455	*89*	0,7%	*76,23*
La Rioja	3.861	*40*	1,0%	*183,71*
ESPAÑA	**207.634**	*1.729*	0,8%	*88,49*

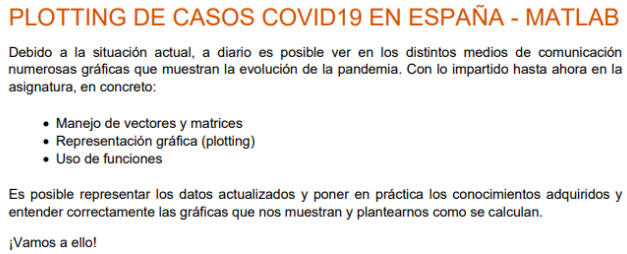
CCAA	Casos que han precisado hospitalización		Casos que han ingresado en UCI		Fallecidos		Curados	
	Total	*Nuevos*	Total	*Nuevos*	Total	*Nuevos*	Total	*Nuevos*
Andalucía	5.748	*33*	716	*0*	1.145	*14*	4.741	*446*
Aragón	2.395	*23*	257	*2*	712	*3*	1.960	*31*
Asturias	1.790	*30*	134	*2*	249	*10*	749	*33*
Baleares	1.062	*4*	166	*0*	175	*1*	1.123	*21*
Canarias	880	*2*	171	*0*	131	*1*	1.048	*12*
Cantabria	991	*9*	78	*0*	183	*1*	1.127	*81*
Castilla La Mancha	8.417	*32*	565	*6*	2.330	*38*	5.196	*320*
Castilla y León	7.653	*98*	515	*4*	1.666	*27*	6.208	*175*
Cataluña	25.665	*1.245*	2.583	*7*	4.566	*68*	17.006	*253*
Ceuta	10	*0*	4	*0*	4	*0*	104	*6*
C. Valenciana	5.013	*35*	654	*1*	1.186	*14*	6.243	*210*
Extremadura	1.480	*33*	108	*1*	422	*6*	1.582	*72*
Galicia	2.735	*13*	85[¥]		394	*6*	1.803	*20*
Madrid	5.892[¥]		873[¥]		7.922	*74*	35.367	*465*
Melilla	44	*0*	3	*0*	2	*0*	81	*4*
Murcia	627	*2*	106	*1*	127	*1*	920	*78*
Navarra	1.942	*5*	129	*0*	431	*2*	1.835	*98*
País Vasco	6.426	*51*	533	*8*	1.230	*18*	9.602	*661*
La Rioja	1.380	*20*	84	*1*	315	*4*	2.037	*38*
ESPAÑA					23.190	*288*	98.732	*3.024*

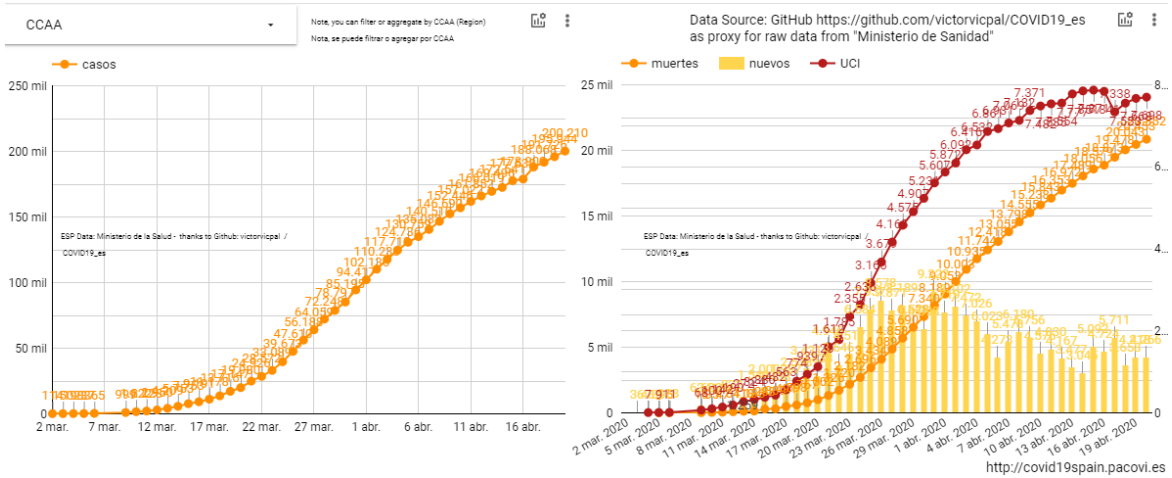
COVID19

@victorvicpal

Things scaled up





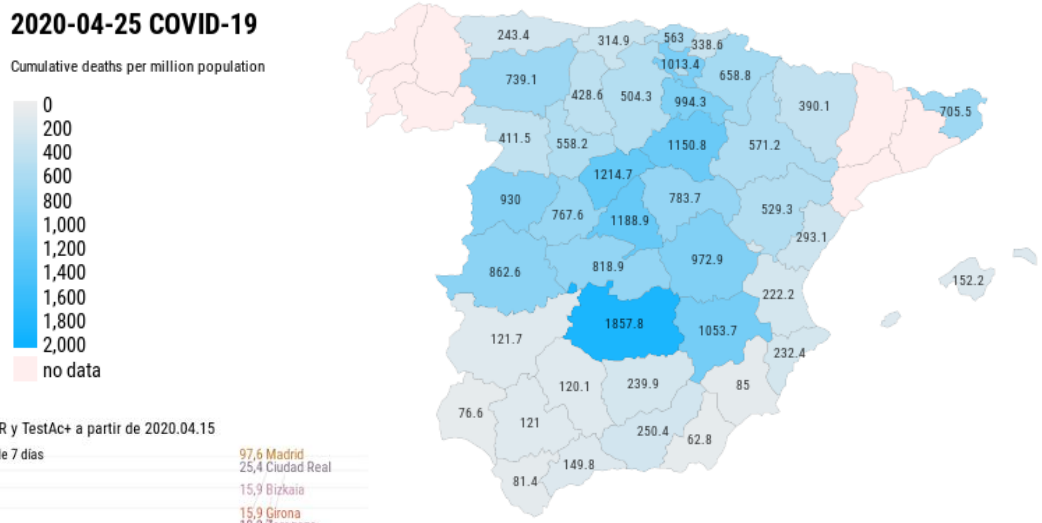


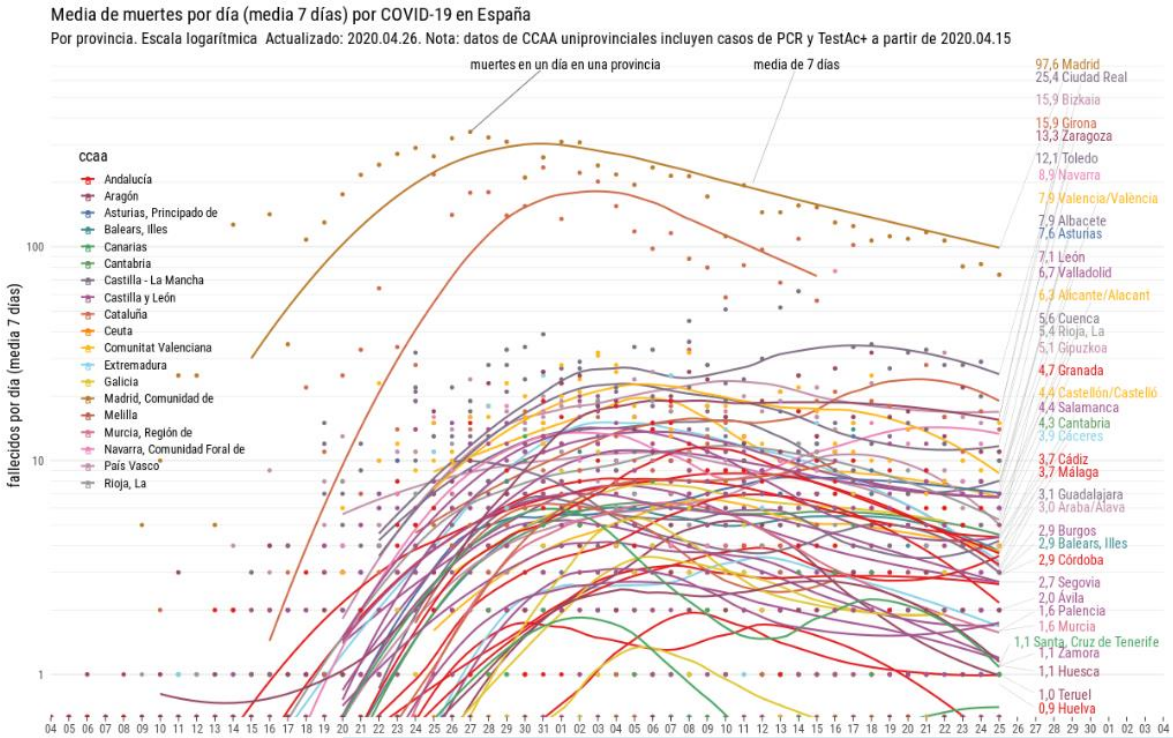
covid-19-stats.info
covid-19-stats

COVID19

@victorvicpal

Numeroteca Datadista



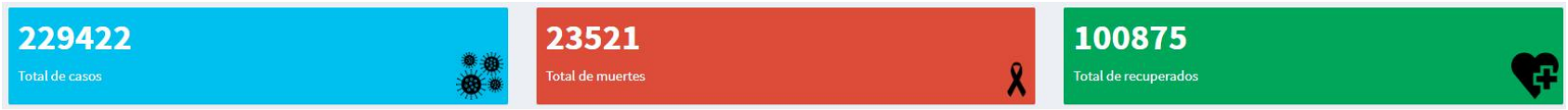


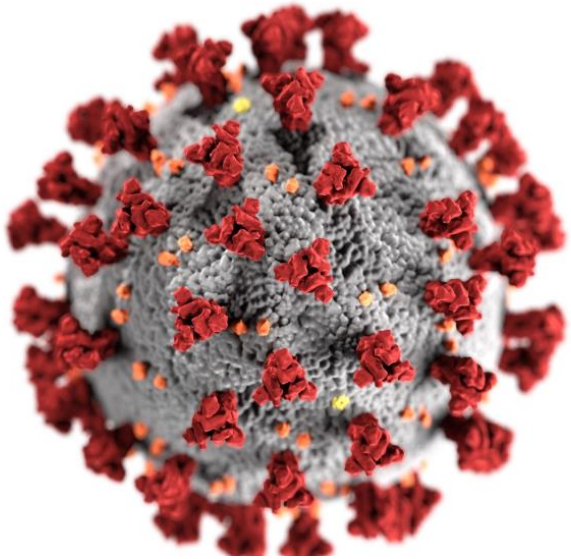
No somos especialistas en epidemiología. Nuestra intención al publicar esta página es poner a disposición gráficos lo más claros posible para proporcionar herramientas que nos ayuden a entender la situación.

COVID19

@victorvicpal

COVID19-Tracker





COVID-19
TRACKER

Una aplicación Shiny para una visualización completa de datos para la epidemia de SARS-CoV-2 en España.

Una aplicació Shiny per una completa visualització de dades de l'epidemia de SARS-CoV-2 a Espanya.

A shiny app to produce to produce comprehensive data visualization for SARS-CoV-2 epidemic in Spain.

https://ubidi.shinyapps.io/covid19/

COVID19

@victorvicpal

How much we might trust COVID19 data?

- The number of new cases each day > poor reflection
- The number of new deaths each day > poor reflection
- The total number of deaths > hopeless tool

MOVING AVERAGE

- Numbers recorded on a logarithmic scale > Just for trends
- Predictions from models > always acknowledge uncertainty
- “Excess deaths” > skepticism
- The lethal risks of being infected > vary with age and frailty
- The “accuracy” of an antibody test > depends

COVID19

@victorvicpal

Questions?