DATA SCIENCE FOR SOCIAL GOOD
PORTUGAL

Data Science for Social Good

Curating a COVID-19 data repository: tales and technicalities

DSPT Meetup #77
The Internet, 7th May 2020

🧪 The formula for Data Science for Social Good

Beneficiaries

Charities
NGOs
Public
Administration

+

Volunteers

Data Scientists
Data Enthusiasts
Developers
Designers
Domain experts

=

Projects

1-3 months; **Deliverable** (small, simple yet able to bring value - report, interactive visualization, predictive model, etc)

⏯ Current status

Rotaract
Santo Tirso

LIGA PORTUGUESA
CONTRA O CANCRO

Past

Associação Zoófila Portuguesa

Currently

VOST.PT

Fruta Feia

CAIS

Soon

A COVID-19 data repository: motivation

Since the start of the pandemic, the Portuguese Health Authorities [Direção Geral da Saúde] has been providing **daily status reports** [1] in a PDF file:

- **Closed** format (not easy to extract knowledge from)
- Unstructured
- No **metadata** or data **dictionaries**
- Highly variable

In parallel, the data community was pumping out lots of analyses. By mid-March, still a **low signal-to-noise ratio** and **everyone was individually scraping the same data**.

[1] https://covid19.min-saude.pt/relatorio-de-situacao/

Motivation: why open and reliable data, anyway?

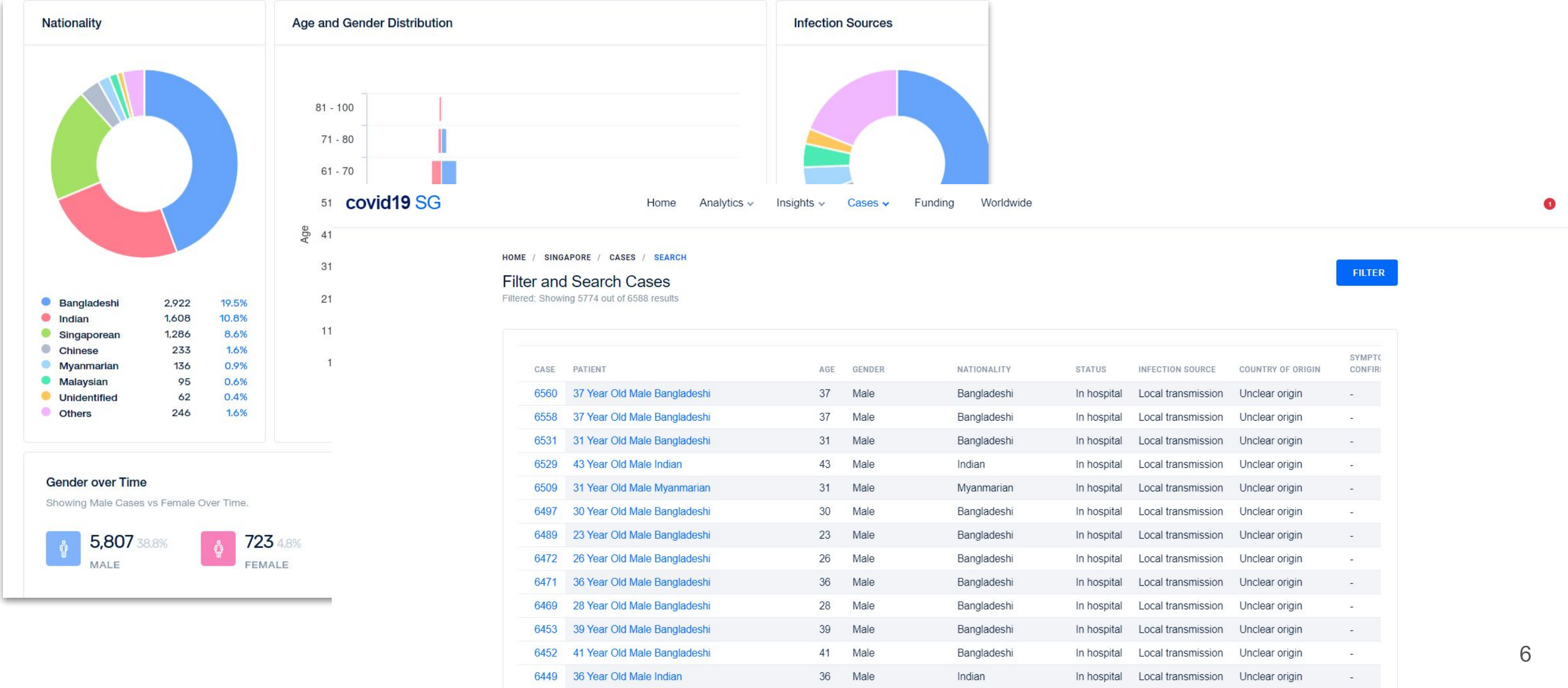
- **Transparency** and **accountability**
- Common **official knowledge base**
- Potential for **civic/industrial/academic** involvement

Two **prime** examples of what open and reliable pandemic data enables:

- **COVID-19 Singapore Dashboard | UCA** [1]
- **GitHub Repository of the Italian Civil Protection** [2]

[1] https://againstcovid19.com/singapore | **[2]** https://github.com/pcm-dpc/COVID-19/

Example: The Singapore dashboard.



● Bangladeshi	2,922	19.5%
● Indian	1,608	10.8%
● Singaporean	1,286	8.6%
● Chinese	233	1.6%
● Myanmarian	136	0.9%
● Malaysian	95	0.6%
● Unidentified	62	0.4%
● Others	246	1.6%

Gender over Time

Showing Male Cases vs Female Over Time.

5,807 38.8% MALE

723 4.8% FEMALE

HOME / SINGAPORE / CASES / SEARCH

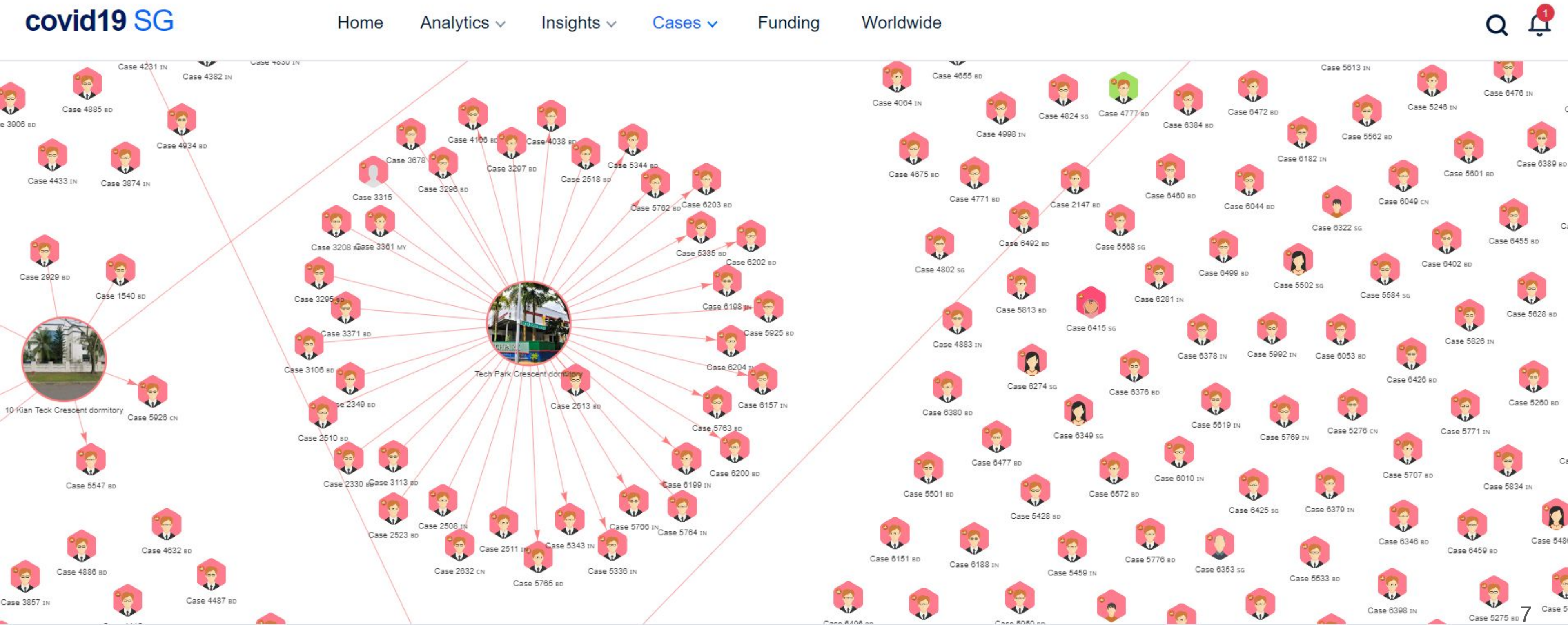
Filter and Search Cases

Filtered: Showing 5774 out of 6588 results

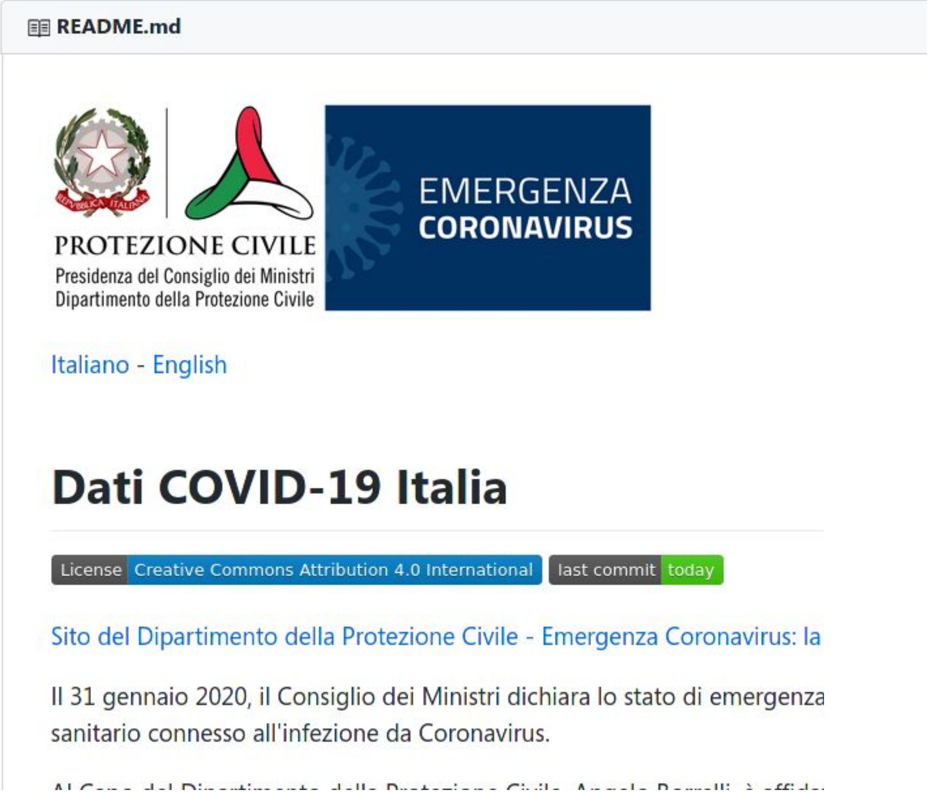
FILTER

CASE	PATIENT	AGE	GENDER	NATIONALITY	STATUS	INFECTION SOURCE	COUNTRY OF ORIGIN	SYMPTC CONFIR
6560	37 Year Old Male Bangladeshi	37	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6558	37 Year Old Male Bangladeshi	37	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6531	31 Year Old Male Bangladeshi	31	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6529	43 Year Old Male Indian	43	Male	Indian	In hospital	Local transmission	Unclear origin	-
6509	31 Year Old Male Myanmarian	31	Male	Myanmarian	In hospital	Local transmission	Unclear origin	-
6497	30 Year Old Male Bangladeshi	30	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6489	23 Year Old Male Bangladeshi	23	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6472	26 Year Old Male Bangladeshi	26	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6471	36 Year Old Male Bangladeshi	36	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6469	28 Year Old Male Bangladeshi	28	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6453	39 Year Old Male Bangladeshi	39	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6452	41 Year Old Male Bangladeshi	41	Male	Bangladeshi	In hospital	Local transmission	Unclear origin	-
6449	36 Year Old Male Indian	36	Male	Indian	In hospital	Local transmission	Unclear origin	-

Example: The Singapore dashboard.



Example: *Italian Civil Protection* data repository



README.md

PROTEZIONE CIVILE
Presidenza del Consiglio dei Ministri
Dipartimento della Protezione Civile

EMERGENZA
CORONAVIRUS

Italiano - English

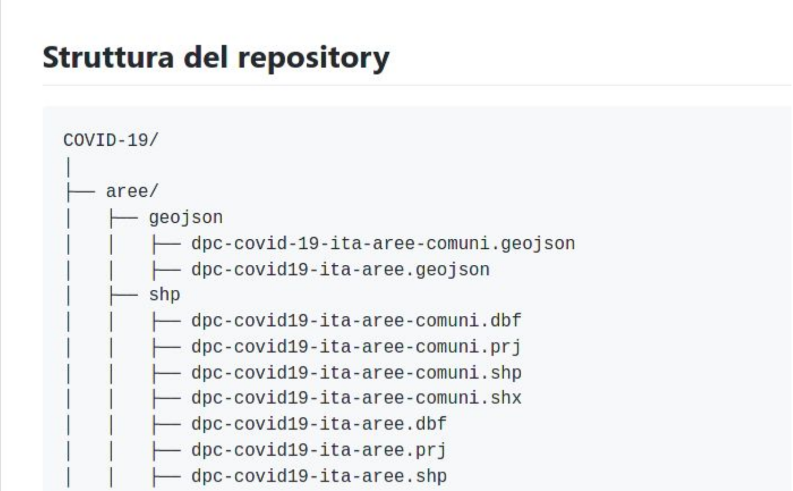
Dati COVID-19 Italia

License | Creative Commons Attribution 4.0 International   last commit | today

Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la

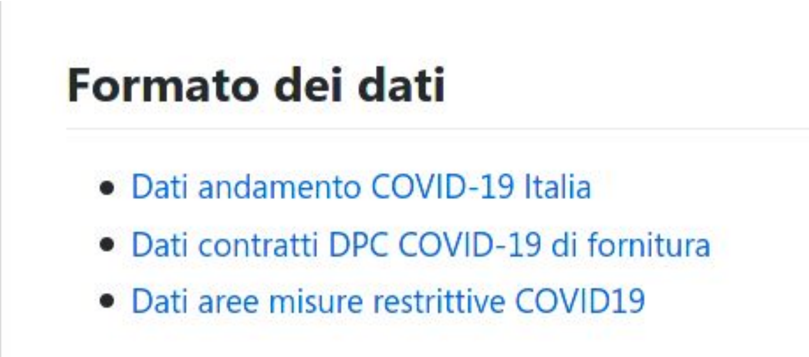
Il 31 gennaio 2020, il Consiglio dei Ministri dichiara lo stato di emergenza sanitario connesso all'infezione da Coronavirus.

Al Capo del Dipartimento della Protezione Civile, Angelo Borrelli, è affida



Struttura del repository

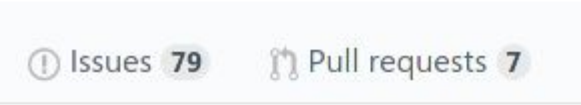
```
COVID-19/
|
├── aree/
|   ├── geojson
|   |   ├── dpc-covid-19-ita-aree-comuni.geojson
|   |   ├── dpc-covid19-ita-aree.geojson
|   ├── shp
|   |   ├── dpc-covid19-ita-aree-comuni.dbf
|   |   ├── dpc-covid19-ita-aree-comuni.prj
|   |   ├── dpc-covid19-ita-aree-comuni.shp
|   |   ├── dpc-covid19-ita-aree-comuni.shx
|   |   ├── dpc-covid19-ita-aree.dbf
|   |   ├── dpc-covid19-ita-aree.prj
|   |   ├── dpc-covid19-ita-aree.shp
```



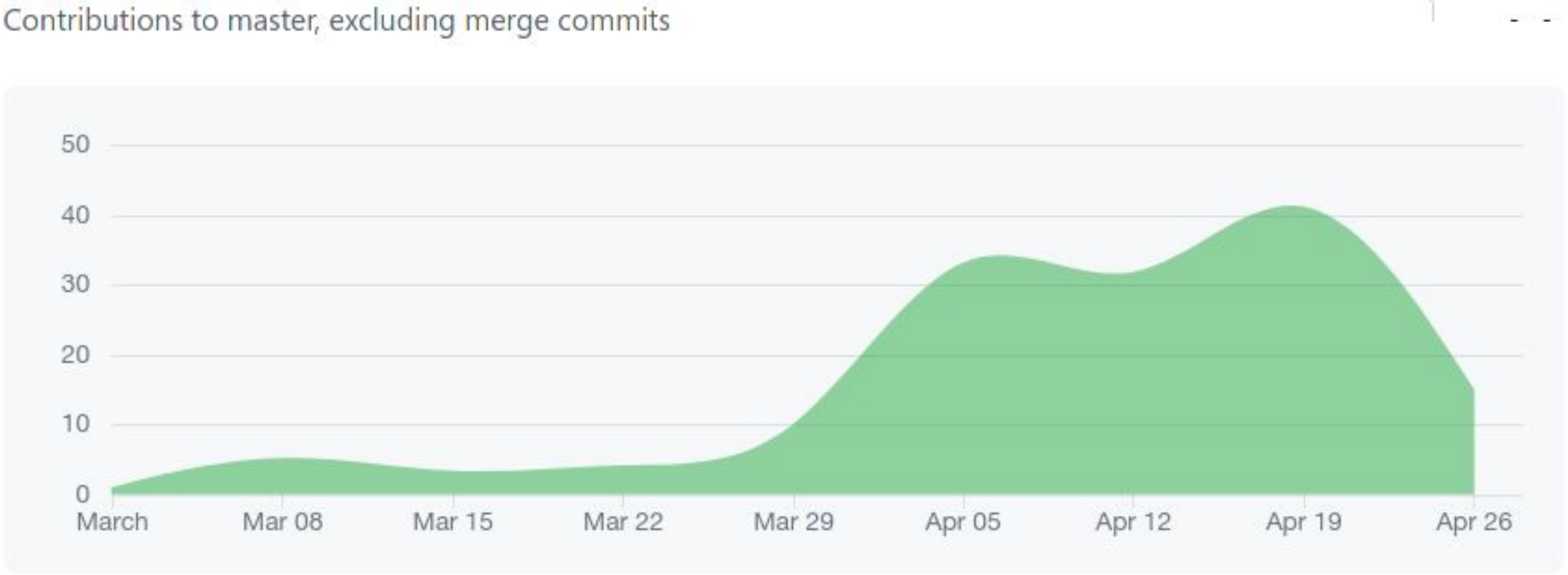
Formato dei dati

- Dati andamento COVID-19 Italia
- Dati contratti DPC COVID-19 di fornitura
- Dati aree misure restrittive COVID19

Example: *Italian Civil Protection* data repository



ⓘ Issues 79 Pull requests 7



Contributions to master, excluding merge commits

🦠 Motivation

👎 No signs of a **concerted data strategy** from the Portuguese Health Authorities

🍕 Can we take it upon ourselves to offer a **central data location** emulating *The Italian Job*?

15/03: Repository launch

dssg-pt / **covid19pt-data**

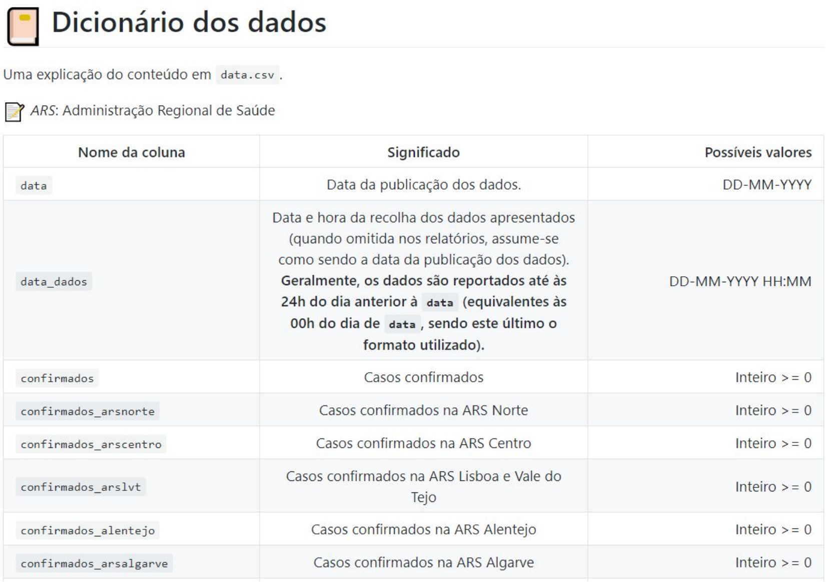
😷🇵🇹 Dados relativos à pandemia COVID-19 em Portugal

Turns out **we can! [1]**

- Started with a **CSV file with epidemiological data** from 26/02 to 14/03 + **archive of the reports**
- At the time, we were manually inserting data but planning on **introducing OCR (Optical Character Recognition)**.
- Had a **comprehensive data dictionary** (first, to our best knowledge).

[1] https://github.com/dssg-pt/covid19pt-dgs-data

15/03: Repository launch



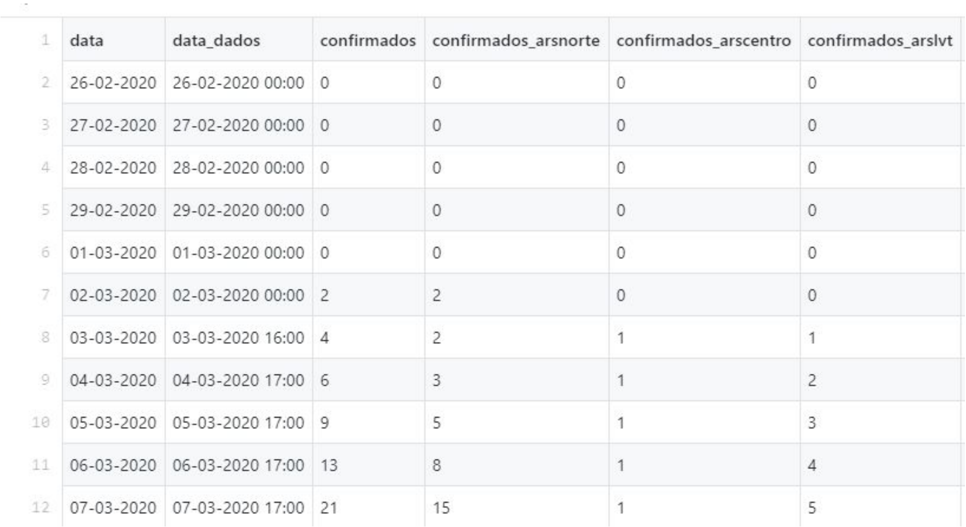
📙 Dicionário dos dados

Uma explicação do conteúdo em `data.csv`.

📝 *ARS*: Administração Regional de Saúde

Nome da coluna	Significado	Possíveis valores
`data`	Data da publicação dos dados.	DD-MM-YYYY
`data_dados`	Data e hora da recolha dos dados apresentados (quando omitida nos relatórios, assume-se como sendo a data da publicação dos dados). **Geralmente, os dados são reportados até às 24h do dia anterior à `data` (equivalentes às 00h do dia de `data`, sendo este último o formato utilizado).**	DD-MM-YYYY HH:MM
`confirmados`	Casos confirmados	Inteiro > = 0
`confirmados_arsnorte`	Casos confirmados na ARS Norte	Inteiro > = 0
`confirmados_arscentro`	Casos confirmados na ARS Centro	Inteiro > = 0
`confirmados_arslvt`	Casos confirmados na ARS Lisboa e Vale do Tejo	Inteiro > = 0
`confirmados_alentejo`	Casos confirmados na ARS Alentejo	Inteiro > = 0
`confirmados_arsalgarve`	Casos confirmados na ARS Algarve	Inteiro > = 0

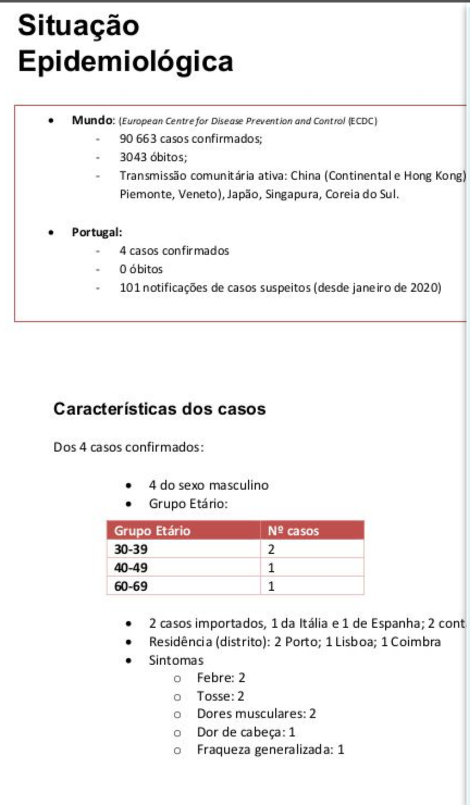
Data Dictionary



data	data_dados	confirmados	confirmados_arsnorte	confirmados_arscentro	confirmados_arslvt
26-02-2020	26-02-2020 00:00	0	0	0	0
27-02-2020	27-02-2020 00:00	0	0	0	0
28-02-2020	28-02-2020 00:00	0	0	0	0
29-02-2020	29-02-2020 00:00	0	0	0	0
01-03-2020	01-03-2020 00:00	0	0	0	0
02-03-2020	02-03-2020 00:00	2	2	0	0
03-03-2020	03-03-2020 16:00	4	2	1	1
04-03-2020	04-03-2020 17:00	6	3	1	2
05-03-2020	05-03-2020 17:00	9	5	1	3
06-03-2020	06-03-2020 17:00	13	8	1	4
07-03-2020	07-03-2020 17:00	21	15	1	5

Data file

OCR would not be a good idea.



Situação Epidemiológica

- **Mundo**: (*European Centre for Disease Prevention and Control* (ECDC)
  - 90 663 casos confirmados;
  - 3043 óbitos;
  - Transmissão comunitária ativa: China (Continental e Hong Kong) Piemonte, Veneto), Japão, Singapura, Coreia do Sul.
- **Portugal:**
  - 4 casos confirmados
  - 0 óbitos
  - 101 notificações de casos suspeitos (desde janeiro de 2020)

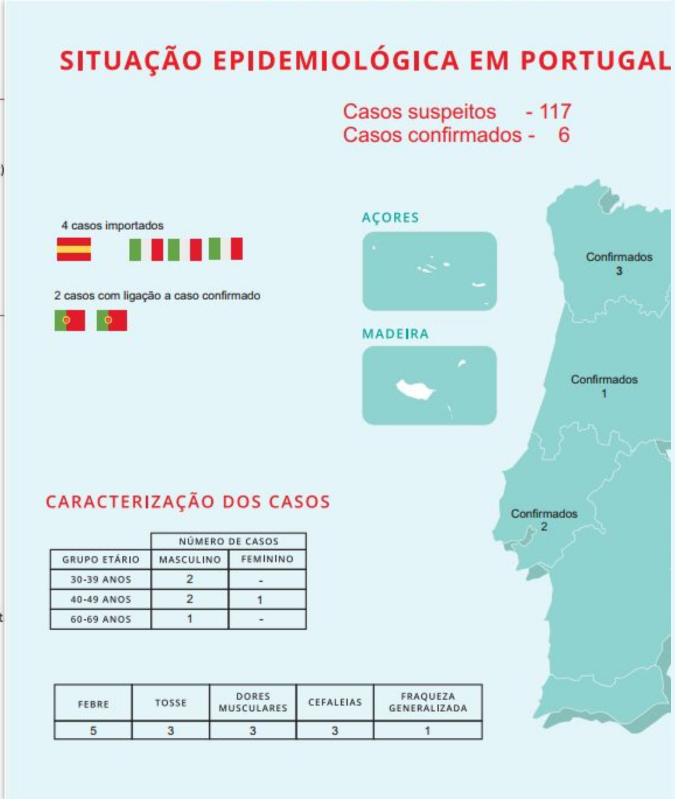
Características dos casos

Dos 4 casos confirmados:

- 4 do sexo masculino
- Grupo Etário:

Grupo Etário	Nº casos
30-39	2
40-49	1
60-69	1

- 2 casos importados, 1 da Itália e 1 de Espanha; 2 cont
- Residência (distrito): 2 Porto; 1 Lisboa; 1 Coimbra
- Sintomas
  - Febre: 2
  - Tosse: 2
  - Dores musculares: 2
  - Dor de cabeça: 1
  - Fraqueza generalizada: 1



SITUAÇÃO EPIDEMIOLÓGICA EM PORTUGAL

Casos suspeitos - 117
Casos confirmados - 6

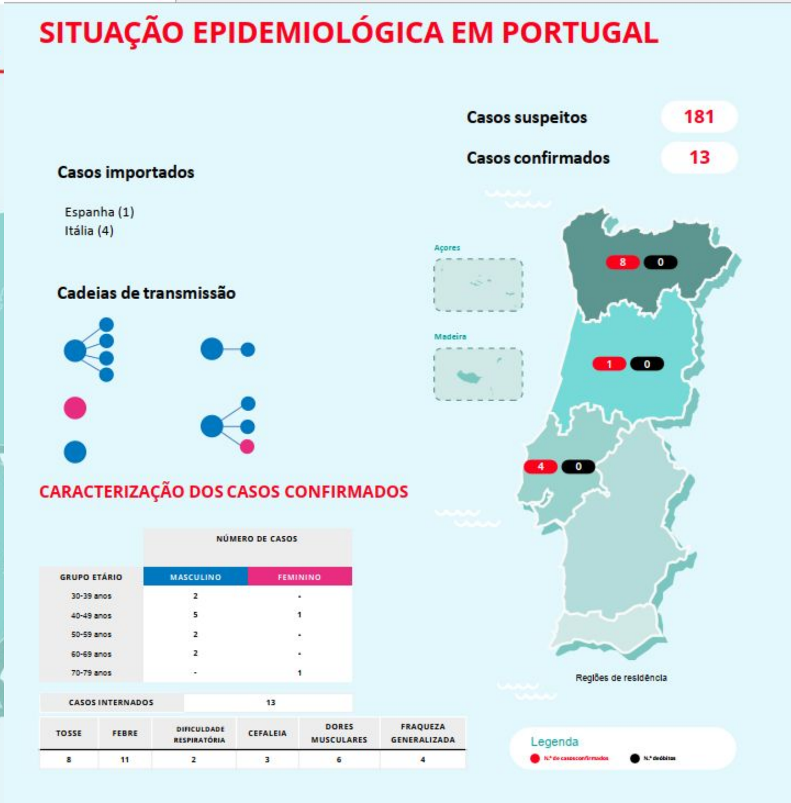
4 casos importados

2 casos com ligação a caso confirmado

CARACTERIZAÇÃO DOS CASOS

GRUPO ETÁRIO	NÚMERO DE CASOS	
	MASCULINO	FEMININO
30-39 ANOS	2	-
40-49 ANOS	2	1
60-69 ANOS	1	-

FEBRE	TOSSE	DORES MUSCULARES	CEFALEIAS	FRAQUEZA GENERALIZADA
5	3	3	3	1



SITUAÇÃO EPIDEMIOLÓGICA EM PORTUGAL

Casos suspeitos	181
Casos confirmados	13

Casos importados

Espanha (1)
Itália (4)

Cadeias de transmissão

CARACTERIZAÇÃO DOS CASOS CONFIRMADOS

GRUPO ETÁRIO	NÚMERO DE CASOS	
	MASCULINO	FEMININO
30-39 anos	2	-
40-49 anos	5	1
50-59 anos	2	-
60-69 anos	2	-
70-79 anos	-	1

CASOS INTERNADOS	13

TOSSE	FEBRE	DIFICULDADE RESPIRATÓRIA	CEFALEIA	DORES MUSCULARES	FRAQUEZA GENERALIZADA
8	11	2	3	6	4

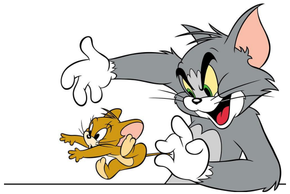
Report #1: 3rd March

Report #2: 4th March

Report #4: 6th March

And so it begins: the cycle of manual updates.

aka **maintaining a data pipeline with no data contract** whatsoever.



And here is **what happened** during that time.

(well, some of it)

🤔 16/03: First extra data sources added

Besides the official data, we have been adding additional **clean data sources/ extra resources** [1] to study the impact of COVID-19 in related areas. As of May 2020:

- **National Healthcare System** (Portal Transparência SNS24)
- **Events** (Civil Protection, Containment Measures)
- **Demographic** (PORDATA)
- **Information Management** (News + Technical Questions [2])
- **Cartographic** (GeoJSON, Shapefiles) [3]

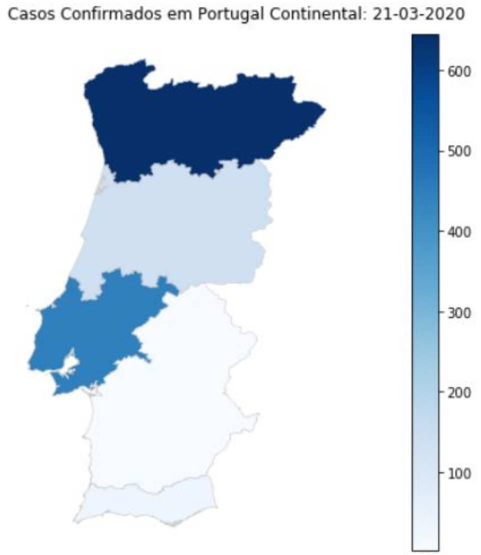
[1] https://github.com/dssg-pt/covid19pt-data/tree/master/extra
[2] Gabriel Mourão, https://github.com/AnthraxisBR
[3] João Palmeiro, https://github.com/joaopalmeiro

🎉 18/03: First community contribution

João Palmeiro added cartographic data (*GeoJSON*, *shapefiles*) for the Portuguese **NUTS II** regions, useful for visualizing COVID-19 cases per region.

Epidemiological data, updated daily

Cartographic data



Pretty and useful cartographic plots

📡 22/03: External API

Carlos Matos (Group IFT) **[1]**, a volunteer, contributed with an **external API built with RapidAPI** **[2]**, for a cleaner access to the epidemiological data in the repository.

It has in the meantime been deprecated, but with the help of the community it took us 8 days from **first commit -> available API**.

[1] https://rapidapi.com/gitgrupoift/api/covid-19-dados-abertos
[2] https://grupoift.pt/

👾 22/03: Sponsoring a new category in TAIKAI Challenge

DSPT got us in touch with TAIKAI. The data became the pillar of one category of their online Hackathon: **Coronavirus Analytics by DSSG**. Ongoing!

FIGHT COVID-19 ONLINE HACKATHON

BY TAIKAI

💪 29/03: Betting on reliability: addition of a test suite

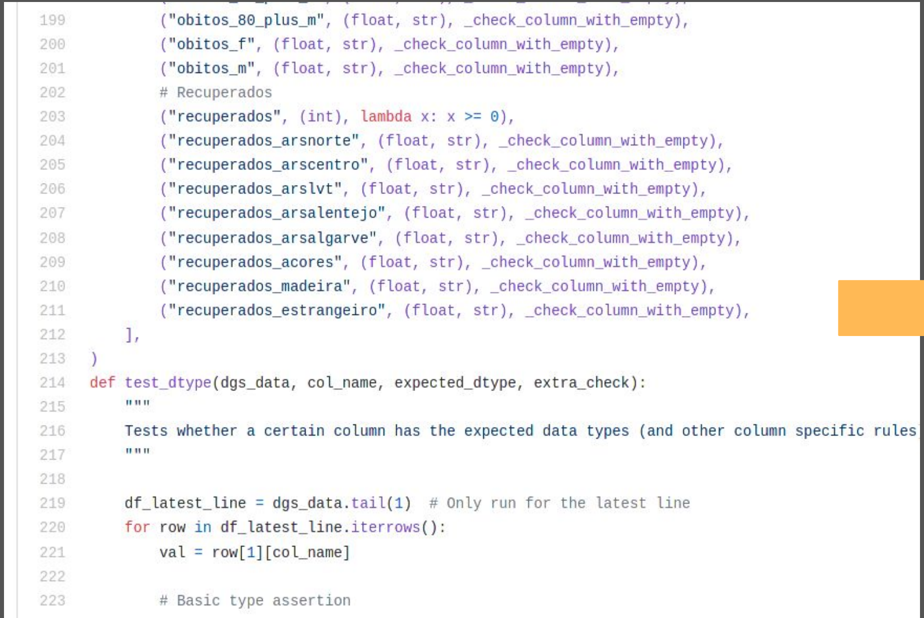
Why?

- As complexity grew, we noticed that the process of manually introducing data was **fallible**.
- Several institutional users were starting to come up - people were using the data for serious things!

Using `pytest`, a Python unit testing library, we developed a suite of **~100 data validations** to guarantee maximum reliability:

- The sum of cases per region must equal the total number of cases
- Columns must have certain data types
- Column separator must be a comma

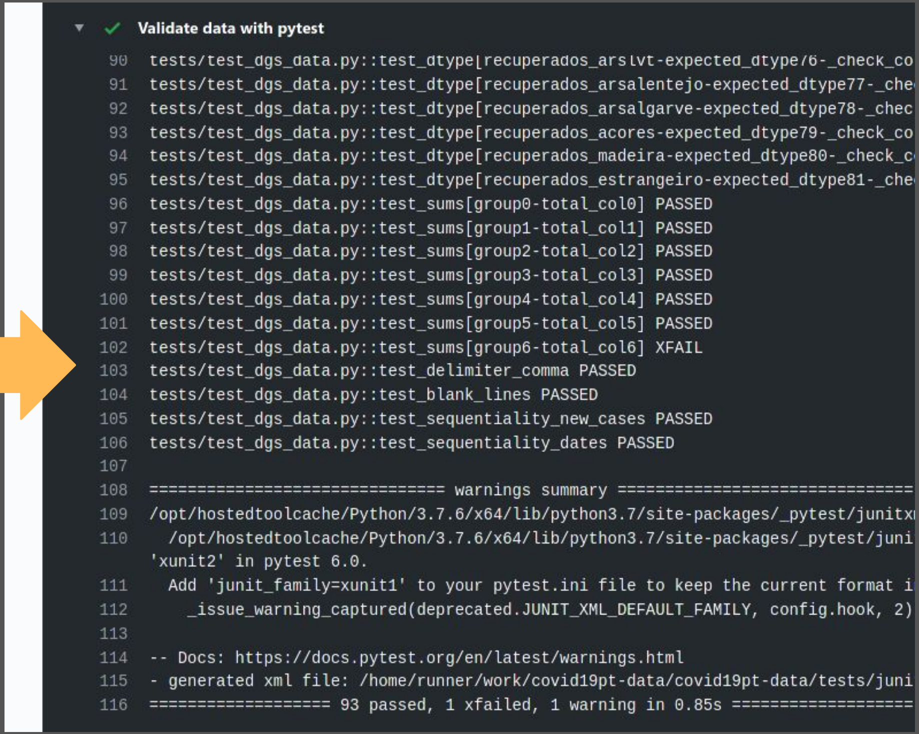
💪 29/03: Betting on reliability: addition of a test suite



```
        ("obitos_80_plus_m", (float, str), _check_column_with_empty),
        ("obitos_f", (float, str), _check_column_with_empty),
        ("obitos_m", (float, str), _check_column_with_empty),
        # Recuperados
        ("recuperados", (int), lambda x: x >= 0),
        ("recuperados_arsnorte", (float, str), _check_column_with_empty),
        ("recuperados_arscentro", (float, str), _check_column_with_empty),
        ("recuperados_arslvt", (float, str), _check_column_with_empty),
        ("recuperados_arsalentejo", (float, str), _check_column_with_empty),
        ("recuperados_arsalgarve", (float, str), _check_column_with_empty),
        ("recuperados_acores", (float, str), _check_column_with_empty),
        ("recuperados_madeira", (float, str), _check_column_with_empty),
        ("recuperados_estrangeiro", (float, str), _check_column_with_empty),
    ],
)
def test_dtype(dgs_data, col_name, expected_dtype, extra_check):
    """
    Tests whether a certain column has the expected data types (and other column specific rules
    """

    df_latest_line = dgs_data.tail(1)  # Only run for the latest line
    for row in df_latest_line.iterrows():
        val = row[1][col_name]

        # Basic type assertion
```



✓ Validate data with pytest

```
tests/test_dgs_data.py::test_dtype[recuperados_arslvt-expected_dtype76-_check_co
tests/test_dgs_data.py::test_dtype[recuperados_arsalentejo-expected_dtype77-_che
tests/test_dgs_data.py::test_dtype[recuperados_arsalgarve-expected_dtype78-_chec
tests/test_dgs_data.py::test_dtype[recuperados_acores-expected_dtype79-_check_co
tests/test_dgs_data.py::test_dtype[recuperados_madeira-expected_dtype80-_check_c
tests/test_dgs_data.py::test_dtype[recuperados_estrangeiro-expected_dtype81-_che
tests/test_dgs_data.py::test_sums[group0-total_col0] PASSED
tests/test_dgs_data.py::test_sums[group1-total_col1] PASSED
tests/test_dgs_data.py::test_sums[group2-total_col2] PASSED
tests/test_dgs_data.py::test_sums[group3-total_col3] PASSED
tests/test_dgs_data.py::test_sums[group4-total_col4] PASSED
tests/test_dgs_data.py::test_sums[group5-total_col5] PASSED
tests/test_dgs_data.py::test_sums[group6-total_col6] XFAIL
tests/test_dgs_data.py::test_delimiter_comma PASSED
tests/test_dgs_data.py::test_blank_lines PASSED
tests/test_dgs_data.py::test_sequentiality_new_cases PASSED
tests/test_dgs_data.py::test_sequentiality_dates PASSED

=============================== warnings summary ===============================
/opt/hostedtoolcache/Python/3.7.6/x64/lib/python3.7/site-packages/_pytest/junitx
  /opt/hostedtoolcache/Python/3.7.6/x64/lib/python3.7/site-packages/_pytest/juni
'xunit2' in pytest 6.0.
  Add 'junit_family=xunit1' to your pytest.ini file to keep the current format i
    _issue_warning_captured(deprecated.JUNIT_XML_DEFAULT_FAMILY, config.hook, 2)

-- Docs: https://docs.pytest.org/en/latest/warnings.html
- generated xml file: /home/runner/work/covid19pt-data/covid19pt-data/tests/juni
=================== 93 passed, 1 xfailed, 1 warning in 0.85s ===================
```

Daily data update: all tests **must** pass + human approval

🗣 9/04: Increasing our initiative’s impact

- By this time, we were updating daily data **for almost a month** and knew the dataset and its shortcomings pretty well.

- **Errors and inconsistencies** from the Portuguese Health Authorities in the daily reports **were at an all-time high.**

- **Public pressure** concerning release of **open pandemic data** was increasing, yet no external signs towards a proper data strategy.

We thought we could help.

🗣️ 9/04: Open Letter to the Portuguese Health Authorities

Gathered a few partner entities for added institutional credibility and medical expertise, namely:

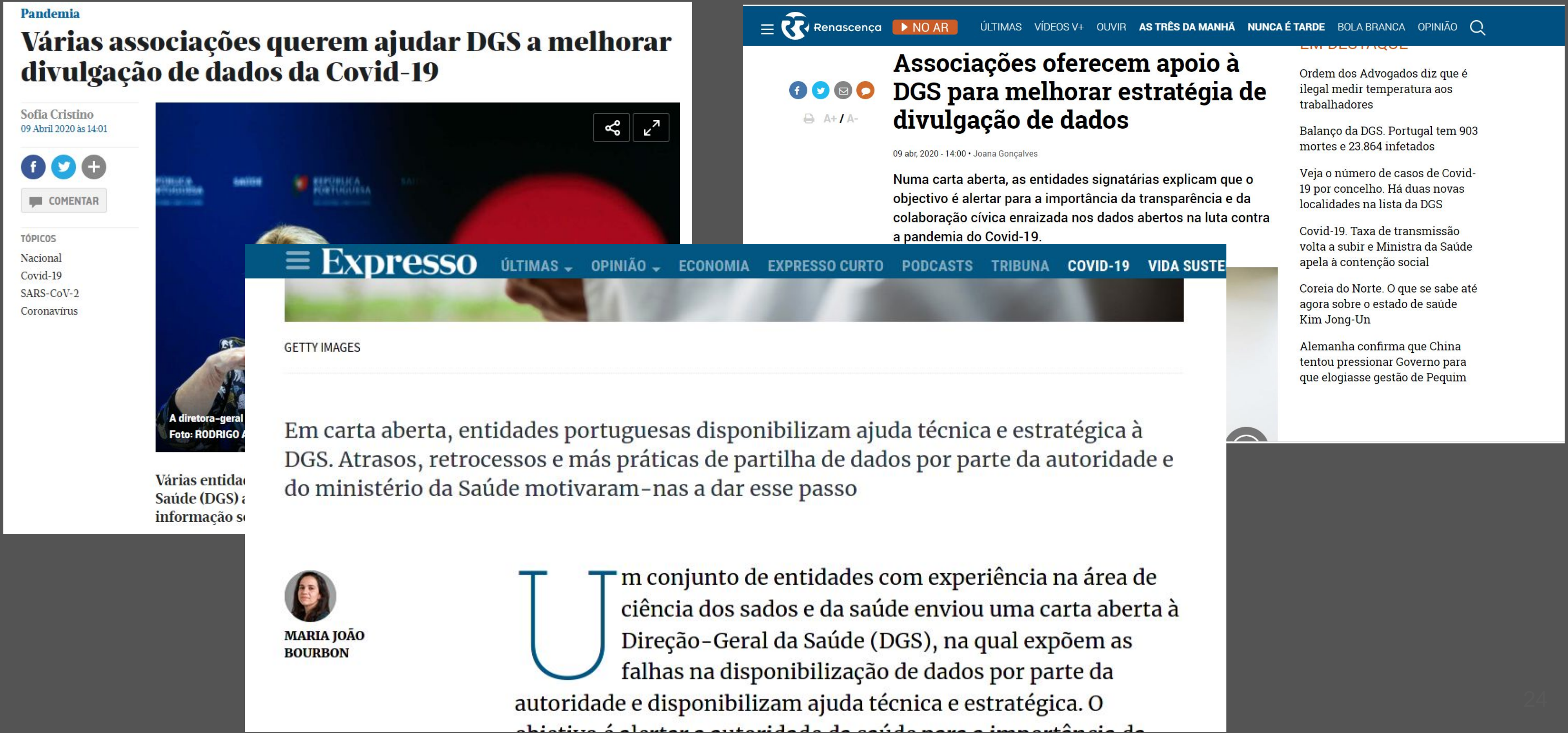
- **Associação Nacional de Médicos de Saúde Pública**
- **Centro de Investigação em Tecnologias e Serviços de Saúde**
- **Faculdade de Medicina da Universidade do Porto**
- **Tech4Covid19 Movement**

Wrote **"Por uma melhor estratégia de dados da Direcção-Geral da Saúde no combate à pandemia COVID-19 em Portugal" [1]**

With constructive and actionable feedback as well as an offer of *pro bono* help, sent to the Head of the Portuguese Health Authorities and, afterwards, Portuguese media.

[1] https://tiny.cc/dgs-openletter

And off it went.



Pandemia

Várias associações querem ajudar DGS a melhorar divulgação de dados da Covid-19

Sofia Cristino
09 Abril 2020 às 14:01

COMENTAR

TÓPICOS
Nacional
Covid-19
SARS-CoV-2
Coronavírus

A diretora-geral
Foto: RODRIGO A

Várias entida
Saúde (DGS) a
informação s

Associações oferecem apoio à DGS para melhorar estratégia de divulgação de dados

A+ / A-

09 abr, 2020 - 14:00 • Joana Gonçalves

Numa carta aberta, as entidades signatárias explicam que o objectivo é alertar para a importância da transparência e da colaboração cívica enraizada nos dados abertos na luta contra a pandemia do Covid-19.

EM DESTAQUE

Ordem dos Advogados diz que é ilegal medir temperatura aos trabalhadores

Balanço da DGS. Portugal tem 903 mortes e 23.864 infetados

Veja o número de casos de Covid-19 por concelho. Há duas novas localidades na lista da DGS

Covid-19. Taxa de transmissão volta a subir e Ministra da Saúde apela à contenção social

Coreia do Norte. O que se sabe até agora sobre o estado de saúde Kim Jong-Un

Alemanha confirma que China tentou pressionar Governo para que elogiasse gestão de Pequim

GETTY IMAGES

Em carta aberta, entidades portuguesas disponibilizam ajuda técnica e estratégica à DGS. Atrasos, retrocessos e más práticas de partilha de dados por parte da autoridade e do ministério da Saúde motivaram-nas a dar esse passo

MARIA JOÃO BOURBON

Um conjunto de entidades com experiência na área de ciência dos sados e da saúde enviou uma carta aberta à Direção-Geral da Saúde (DGS), na qual expõem as falhas na disponibilização de dados por parte da autoridade e disponibilizam ajuda técnica e estratégica. O objetivo é alertar a autoridade da saúde para a importância da

🗣 9/04: Open Letter - End result

- Currently trying to schedule a meeting with the Portuguese Health Authorities.
- **Overwhelming support** in social networks.
- Kept the **open pandemic data issue in the headlines**, further exerting public pressure.
- Led to a **Motion for Resolution** in the Portuguese Parliament (in analysis).

💭 Data, more and more, a cause, consequence and vehicle for civic intervention.

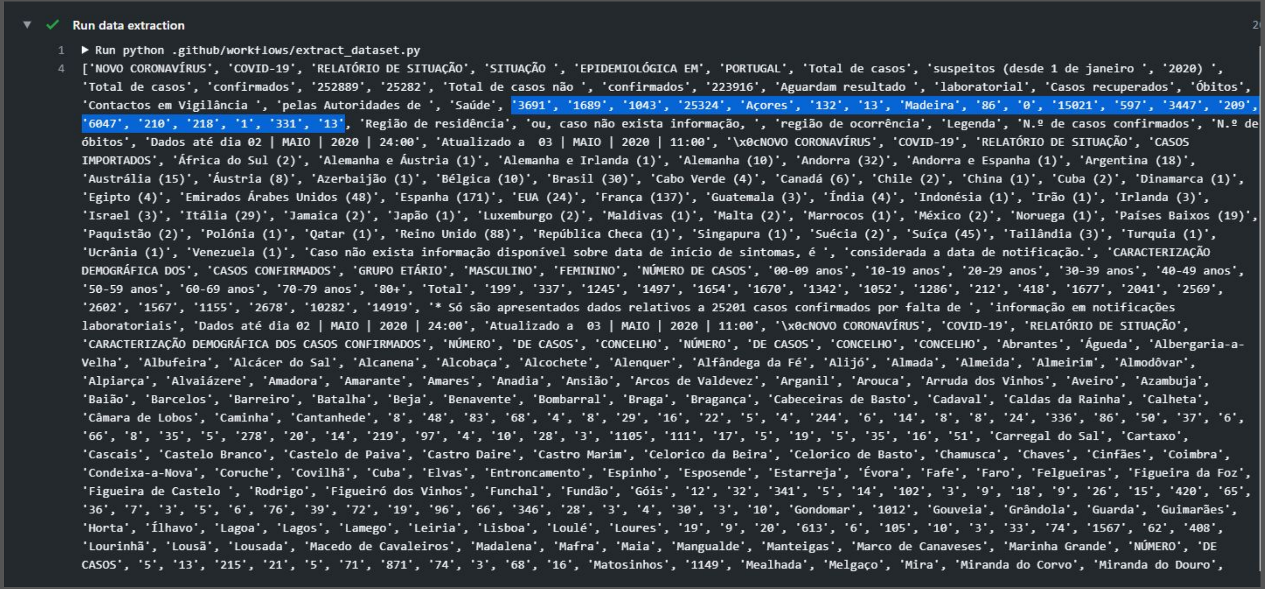
👩‍💻 11/04: Automatic PDF scraping... at last!

Teresa Salazar (from Talkdesk) **[1]** developed a script to automatically extract data from the PDF to our CSV.

- This meant we did not have to manually fill **84 columns** every day.
- Used the `textract` Python library for **text extraction**, followed by heuristic processing rules (doable as the format had stabilized)
- Library is **sensitive configuration of the host OS**: a Docker environment is recommended for reproducibility

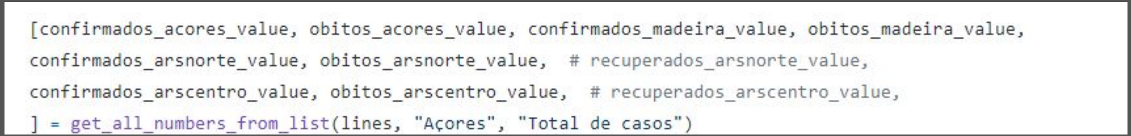
[1] https://github.com/teresalazar13

🤖 Automatic Extraction



Run data extraction

```
▶ Run python .github/workflows/extract_dataset.py
['NOVO CORONAVÍRUS', 'COVID-19', 'RELATÓRIO DE SITUAÇÃO', 'SITUAÇÃO ', 'EPIDEMIOLÓGICA EM', 'PORTUGAL', 'Total de casos', 'suspeitos (desde 1 de janeiro ', '2020) ',
'Total de casos', 'confirmados', '252889', '25282', 'Total de casos não ', 'confirmados', '223916', 'Aguardam resultado ', 'laboratorial', 'Casos recuperados', 'Óbitos',
'Contactos em Vigilância ', 'pelas Autoridades de ', 'Saúde', '3691', '1689', '1043', '25324', 'Açores', '132', '13', 'Madeira', '86', '0', '15021', '597', '3447', '209',
'6047', '210', '218', '1', '331', '13', 'Região de residência', 'ou, caso não exista informação, ', 'região de ocorrência', 'Legenda', 'N.º de casos confirmados', 'N.º de
óbitos', 'Dados até dia 02 | MAIO | 2020 | 24:00', 'Atualizado a  03 | MAIO | 2020 | 11:00', '\x0cNOVO CORONAVÍRUS', 'COVID-19', 'RELATÓRIO DE SITUAÇÃO', 'CASOS
IMPORTADOS', 'África do Sul (2)', 'Alemanha e Áustria (1)', 'Alemanha e Irlanda (1)', 'Alemanha (10)', 'Andorra (32)', 'Andorra e Espanha (1)', 'Argentina (18)',
'Austrália (15)', 'Áustria (8)', 'Azerbaijão (1)', 'Bélgica (10)', 'Brasil (30)', 'Cabo Verde (4)', 'Canadá (6)', 'Chile (2)', 'China (1)', 'Cuba (2)', 'Dinamarca (1)',
'Egipto (4)', 'Emirados Árabes Unidos (48)', 'Espanha (171)', 'EUA (24)', 'França (137)', 'Guatemala (3)', 'Índia (4)', 'Indonésia (1)', 'Irão (1)', 'Irlanda (3)',
'Israel (3)', 'Itália (29)', 'Jamaica (2)', 'Japão (1)', 'Luxemburgo (2)', 'Maldivas (1)', 'Malta (2)', 'Marrocos (1)', 'México (2)', 'Noruega (1)', 'Países Baixos (19)',
'Paquistão (2)', 'Polónia (1)', 'Qatar (1)', 'Reino Unido (88)', 'República Checa (1)', 'Singapura (1)', 'Suécia (2)', 'Suíça (45)', 'Tailândia (3)', 'Turquia (1)',
'Ucrânia (1)', 'Venezuela (1)', 'Caso não exista informação disponível sobre data de início de sintomas, é ', 'considerada a data de notificação.', 'CARACTERIZAÇÃO
DEMOGRÁFICA DOS', 'CASOS CONFIRMADOS', 'GRUPO ETÁRIO', 'MASCULINO', 'FEMININO', 'NÚMERO DE CASOS', '00-09 anos', '10-19 anos', '20-29 anos', '30-39 anos', '40-49 anos',
'50-59 anos', '60-69 anos', '70-79 anos', '80+', 'Total', '199', '337', '1245', '1497', '1654', '1670', '1342', '1052', '1286', '212', '418', '1677', '2041', '2569',
'2602', '1567', '1155', '2678', '10282', '14919', '* Só são apresentados dados relativos a 25201 casos confirmados por falta de ', 'informação em notificações
laboratoriais', 'Dados até dia 02 | MAIO | 2020 | 24:00', 'Atualizado a  03 | MAIO | 2020 | 11:00', '\x0cNOVO CORONAVÍRUS', 'COVID-19', 'RELATÓRIO DE SITUAÇÃO',
'CARACTERIZAÇÃO DEMOGRÁFICA DOS CASOS CONFIRMADOS', 'NÚMERO', 'DE CASOS', 'CONCELHO', 'NÚMERO', 'DE CASOS', 'CONCELHO', 'CONCELHO', 'Abrantes', 'Águeda', 'Albergaria-a-
Velha', 'Albufeira', 'Alcácer do Sal', 'Alcanena', 'Alcobaça', 'Alcochete', 'Alenquer', 'Alfândega da Fé', 'Alijó', 'Almada', 'Almeida', 'Almeirim', 'Almodôvar',
'Alpiarça', 'Alvaiázere', 'Amadora', 'Amarante', 'Amares', 'Anadia', 'Ansião', 'Arcos de Valdevez', 'Arganil', 'Arouca', 'Arruda dos Vinhos', 'Aveiro', 'Azambuja',
'Baião', 'Barcelos', 'Barreiro', 'Batalha', 'Beja', 'Benavente', 'Bombarral', 'Braga', 'Bragança', 'Cabeceiras de Basto', 'Cadaval', 'Caldas da Rainha', 'Calheta',
'Câmara de Lobos', 'Caminha', 'Cantanhede', '8', '48', '83', '68', '4', '8', '29', '16', '22', '5', '4', '244', '6', '14', '8', '8', '24', '336', '86', '50', '37', '6',
'66', '8', '35', '5', '278', '20', '14', '219', '97', '4', '10', '28', '3', '1105', '111', '17', '5', '19', '5', '35', '16', '51', 'Carregal do Sal', 'Cartaxo',
'Cascais', 'Castelo Branco', 'Castelo de Paiva', 'Castro Daire', 'Castro Marim', 'Celorico da Beira', 'Celorico de Basto', 'Chamusca', 'Chaves', 'Cinfães', 'Coimbra',
'Condeixa-a-Nova', 'Coruche', 'Covilhã', 'Cuba', 'Elvas', 'Entroncamento', 'Espinho', 'Esposende', 'Estarreja', 'Évora', 'Fafe', 'Faro', 'Felgueiras', 'Figueira da Foz',
'Figueira de Castelo ', 'Rodrigo', 'Figueiró dos Vinhos', 'Funchal', 'Fundão', 'Góis', '12', '32', '341', '5', '14', '102', '3', '9', '18', '9', '26', '15', '420', '65',
'36', '7', '3', '5', '6', '76', '39', '72', '19', '96', '66', '346', '28', '3', '4', '30', '3', '10', 'Gondomar', '1012', 'Gouveia', 'Grândola', 'Guarda', 'Guimarães',
'Horta', 'Ílhavo', 'Lagoa', 'Lagos', 'Lamego', 'Leiria', 'Lisboa', 'Loulé', 'Loures', '19', '9', '20', '613', '6', '105', '10', '3', '33', '74', '1567', '62', '408',
'Lourinhã', 'Lousã', 'Lousada', 'Macedo de Cavaleiros', 'Madalena', 'Mafra', 'Maia', 'Mangualde', 'Manteigas', 'Marco de Canaveses', 'Marinha Grande', 'NÚMERO', 'DE
CASOS', '5', '13', '215', '21', '5', '71', '871', '74', '3', '68', '16', 'Matosinhos', '1149', 'Mealhada', 'Melgaço', 'Mira', 'Miranda do Corvo', 'Miranda do Douro',
```



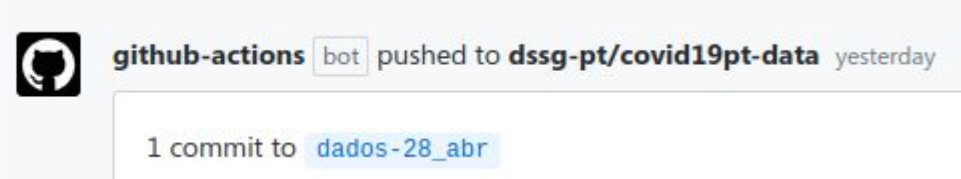
```
[confirmados_acores_value, obitos_acores_value, confirmados_madeira_value, obitos_madeira_value,
confirmados_arsnorte_value, obitos_arsnorte_value,  # recuperados_arsnorte_value,
confirmados_arscentro_value, obitos_arscentro_value,  # recuperados_arscentro_value,
] = get_all_numbers_from_list(lines, "Açores", "Total de casos")
```

🤖 15/04: Fully automated data extraction pipeline

Almost fully automated. We used **Github Actions** for running a data extraction pipeline:

- Pull Request **changing the report link** (exact name varies daily)
- **Trigger the workflow:** download the PDF, extract data, update the CSV and then run the test suite to validate it.
- Before merging, the results are **validated** by a member of our Lead Team.

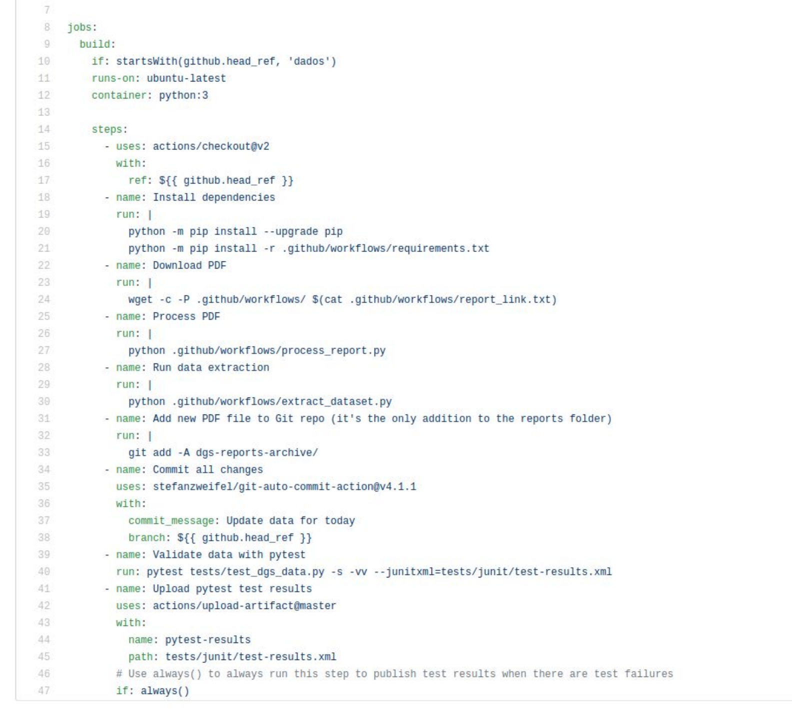
This has been running (somewhat) smoothly **for about two weeks**, with minor changes.



github-actions bot pushed to **dssg-pt/covid19pt-data** yesterday

1 commit to dados-28_abr

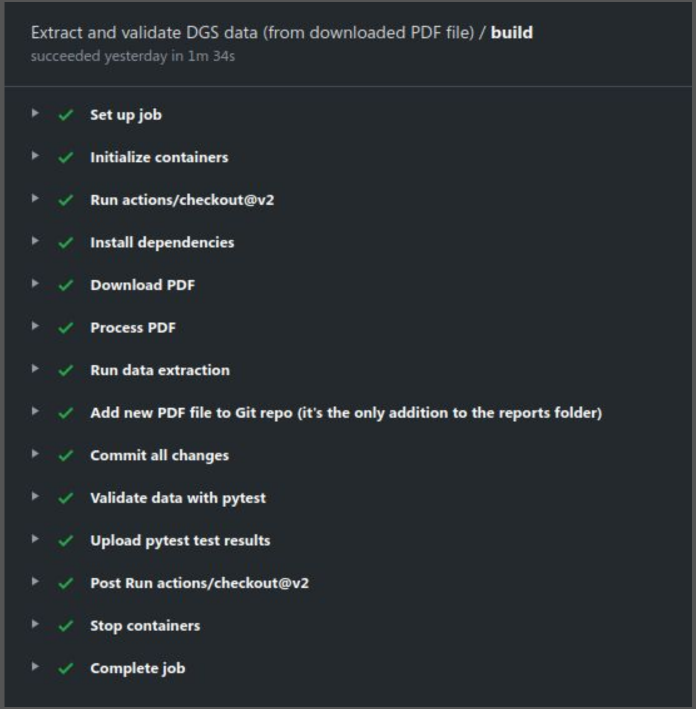
🤖 GitHub Actions?



```yaml

jobs:
  build:
    if: startsWith(github.head_ref, 'dados')
    runs-on: ubuntu-latest
    container: python:3

    steps:
      - uses: actions/checkout@v2
        with:
          ref: ${{ github.head_ref }}
      - name: Install dependencies
        run: |
          python -m pip install --upgrade pip
          python -m pip install -r .github/workflows/requirements.txt
      - name: Download PDF
        run: |
          wget -c -P .github/workflows/ $(cat .github/workflows/report_link.txt)
      - name: Process PDF
        run: |
          python .github/workflows/process_report.py
      - name: Run data extraction
        run: |
          python .github/workflows/extract_dataset.py
      - name: Add new PDF file to Git repo (it's the only addition to the reports folder)
        run: |
          git add -A dgs-reports-archive/
      - name: Commit all changes
        uses: stefanzweifel/git-auto-commit-action@v4.1.1
        with:
          commit_message: Update data for today
          branch: ${{ github.head_ref }}
      - name: Validate data with pytest
        run: pytest tests/test_dgs_data.py -s -vv --junitxml=tests/junit/test-results.xml
      - name: Upload pytest test results
        uses: actions/upload-artifact@master
        with:
          name: pytest-results
          path: tests/junit/test-results.xml
        # Use always() to always run this step to publish test results when there are test failures
        if: always()
```



Extract and validate DGS data (from downloaded PDF file) / **build**

succeeded yesterday in 1m 34s

- ✓ **Set up job**
- ✓ **Initialize containers**
- ✓ **Run actions/checkout@v2**
- ✓ **Install dependencies**
- ✓ **Download PDF**
- ✓ **Process PDF**
- ✓ **Run data extraction**
- ✓ **Add new PDF file to Git repo (it's the only addition to the reports folder)**
- ✓ **Commit all changes**
- ✓ **Validate data with pytest**
- ✓ **Upload pytest test results**
- ✓ **Post Run actions/checkout@v2**
- ✓ **Stop containers**
- ✓ **Complete job**

Free service: human-readable YAML syntax allows definition of **environment** and **arbitrary workflows**; automatic machine provisioning; ideal for **light workloads with specific triggers**.

📡 21/04: API for epidemiological data, take 2

Simple `flask` API (documented with `flask-swagger`) for **access to epidemiological data**, consuming directly from the repository (**always up to date**). It is served using `gunicorn` + `nginx` and deployed using a Docker Container.

- Last update
- Data of a specific day
- Data for a range of days
- Of course, **open-source [1]**. **VOST PT is kindly hosting a public instance [2]**, go make stuff with it!

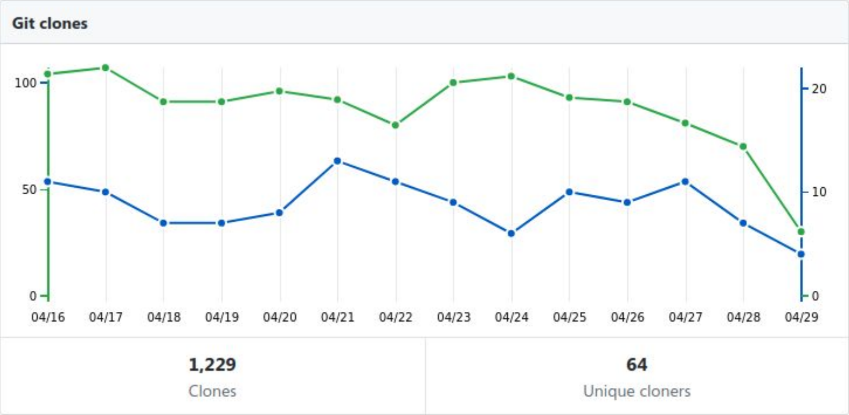
[1] https://github.com/dssg-pt/Docker_COVID_API
[2] https://covid19-api.vost.pt/

📊 Status after 2 months

~300 commits

~60 forks

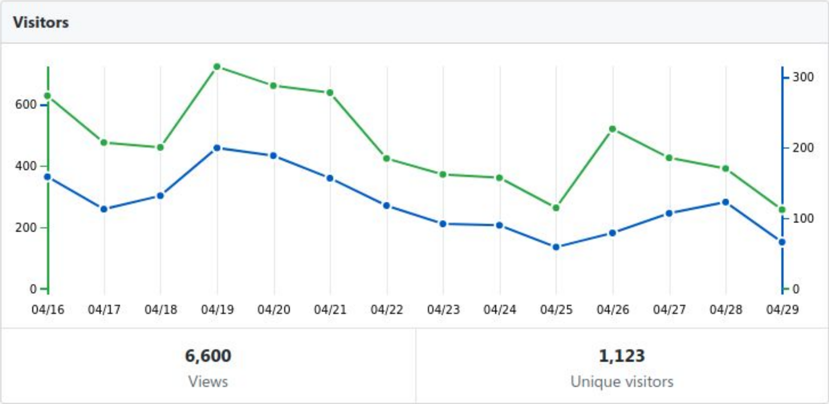
~200 GitHub ⭐



Git clones

1,229
Clones

64
Unique cloners



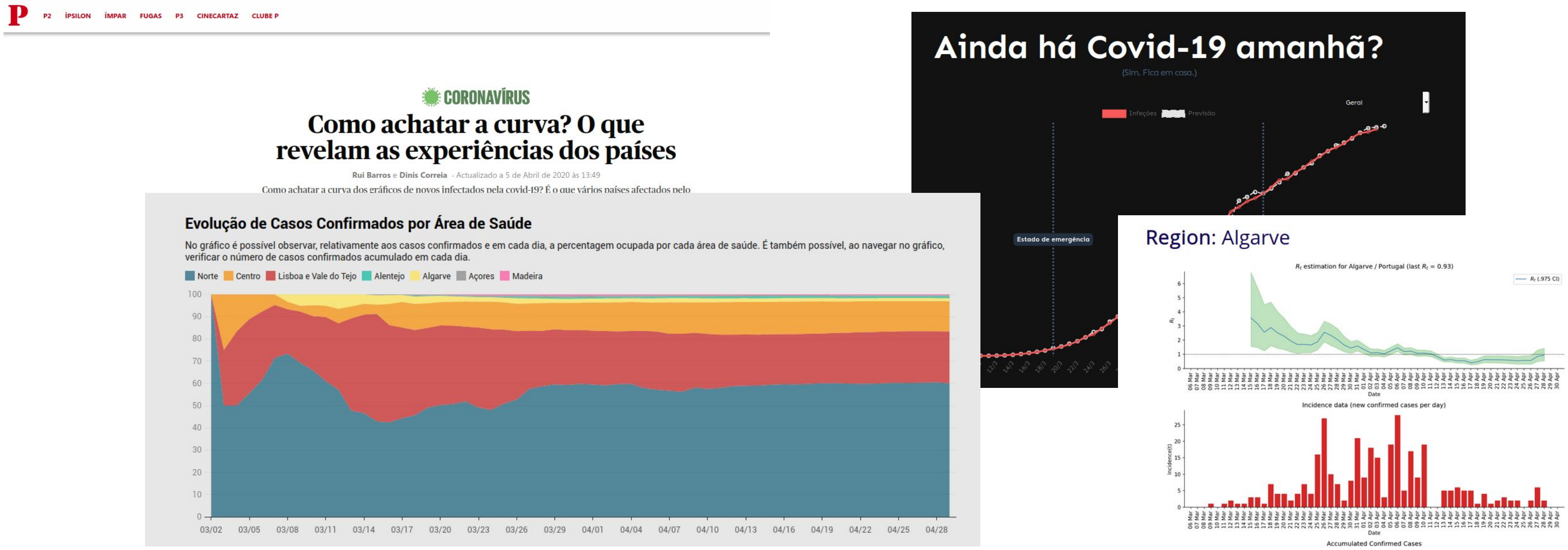
Visitors

6,600
Views

1,123
Unique visitors

👀 Project showcase

`dssg-pt/covid19pt-data/` is powering projects in **data visualization**, **data journalism**, **epidemiological surveys**, **predictive epidemiological modelling**, between others. Projects are personal and academic.



Como achatar a curva? O que revelam as experiências dos países

Rui Barros e Dinis Correia · Actualizado a 5 de Abril de 2020 às 13:49

Como achatar a curva dos gráficos de novos infectados pela covid-19? É o que vários países afectados pelo

Evolução de Casos Confirmados por Área de Saúde

No gráfico é possível observar, relativamente aos casos confirmados e em cada dia, a percentagem ocupada por cada área de saúde. É também possível, ao navegar no gráfico, verificar o número de casos confirmados acumulado em cada dia.

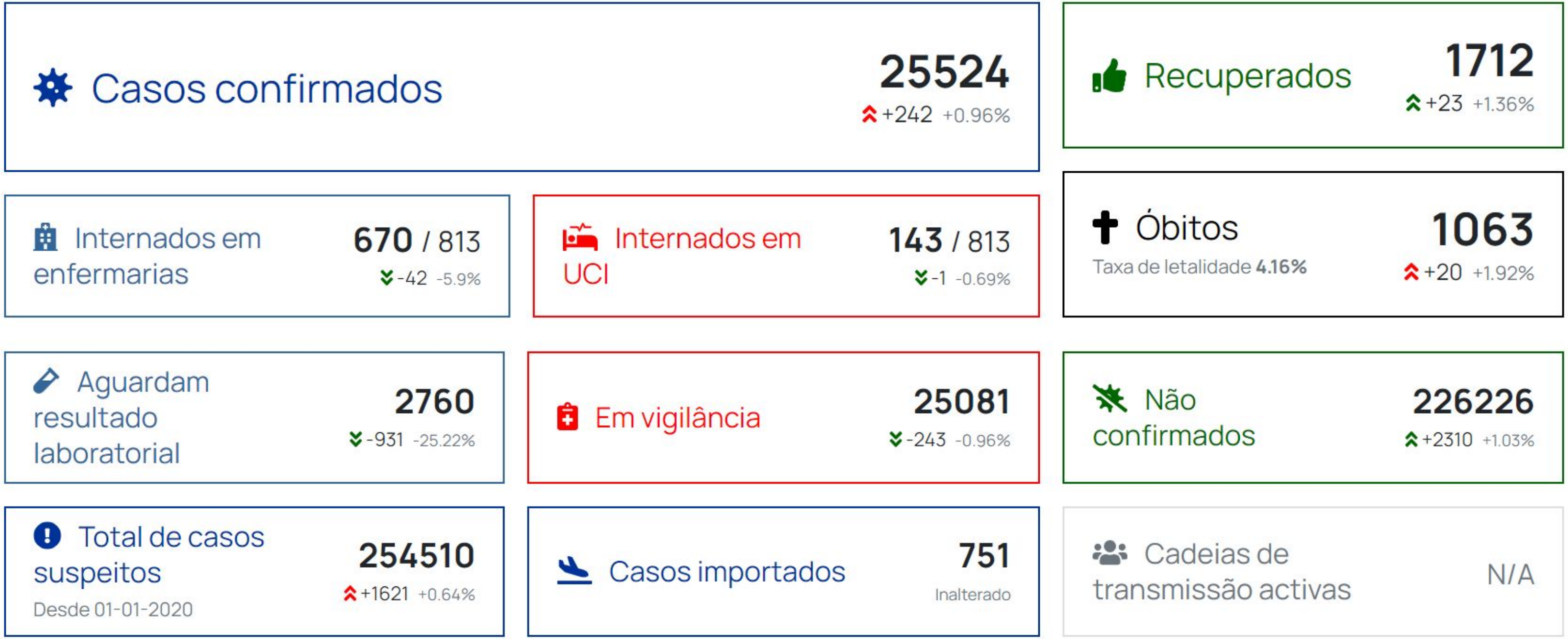
Ainda há Covid-19 amanhã?

(Sim. Fica em casa.)

Region: Algarve

R_t estimation for Algarve / Portugal (last R_t = 0.93)
R_t (.975 CI)
R_t
Date
Incidence data (new confirmed cases per day)
Incidence(t)
Date
Accumulated Confirmed Cases
06 Mar
30 Apr

👀 Antero Pires - COVID19 Dashboard



Casos confirmados 25524
⇈ +242 +0.96%

Recuperados 1712
⇈ +23 +1.36%

Internados em enfermarias 670 / 813
⇊ -42 -5.9%

Internados em UCI 143 / 813
⇊ -1 -0.69%

✝ Óbitos 1063
Taxa de letalidade **4.16%**
⇈ +20 +1.92%

Aguardam resultado laboratorial 2760
⇊ -931 -25.22%

Em vigilância 25081
⇊ -243 -0.96%

Não confirmados 226226
⇈ +2310 +1.03%

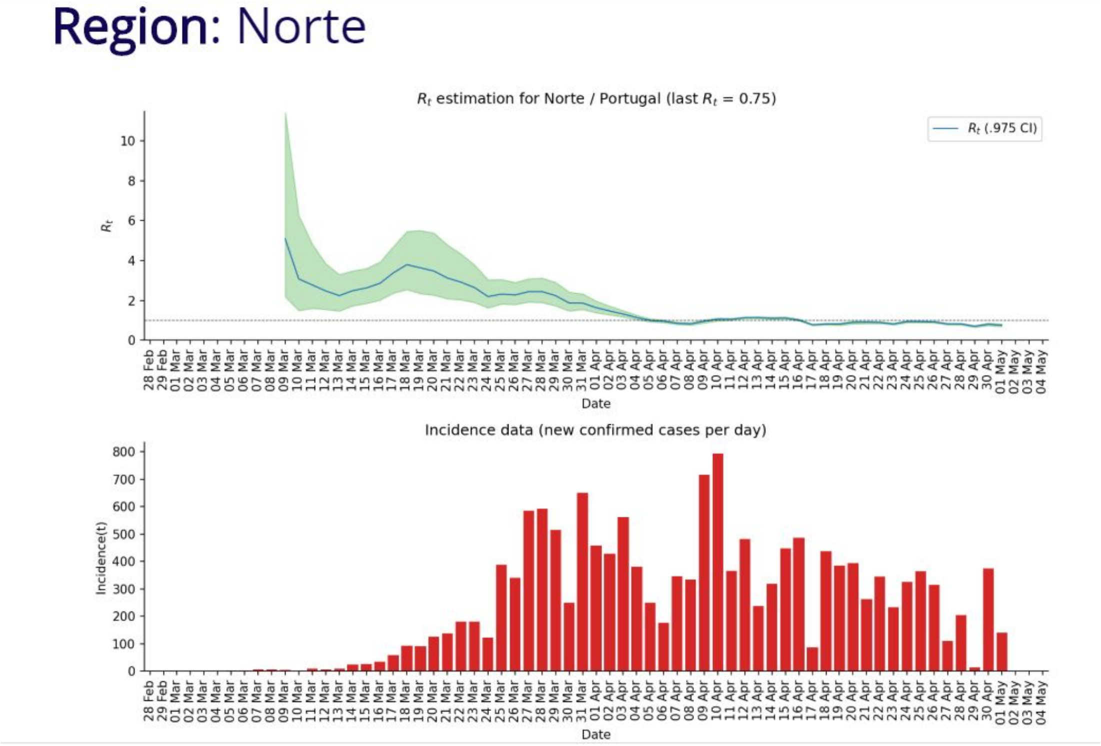
Total de casos suspeitos 254510
Desde 01-01-2020
⇈ +1621 +0.64%

Casos importados 751
Inalterado

Cadeias de transmissão activas N/A

[1] https://covid19.anteropires.com/

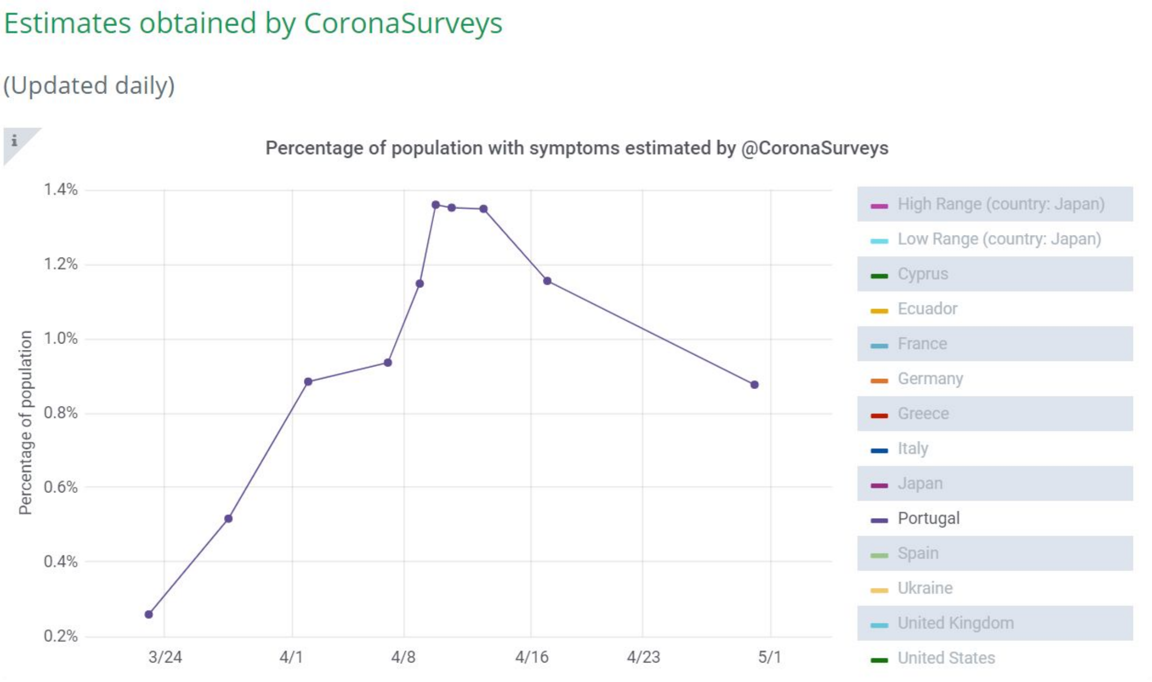
👀 Christian Perone - Rt estimation



Region: Norte

[1] https://perone.github.io/covid19analysis/portugal_r0.html **[2]** https://github.com/perone

👀 Coronasurveys - INESC TEC + Univ. Minho

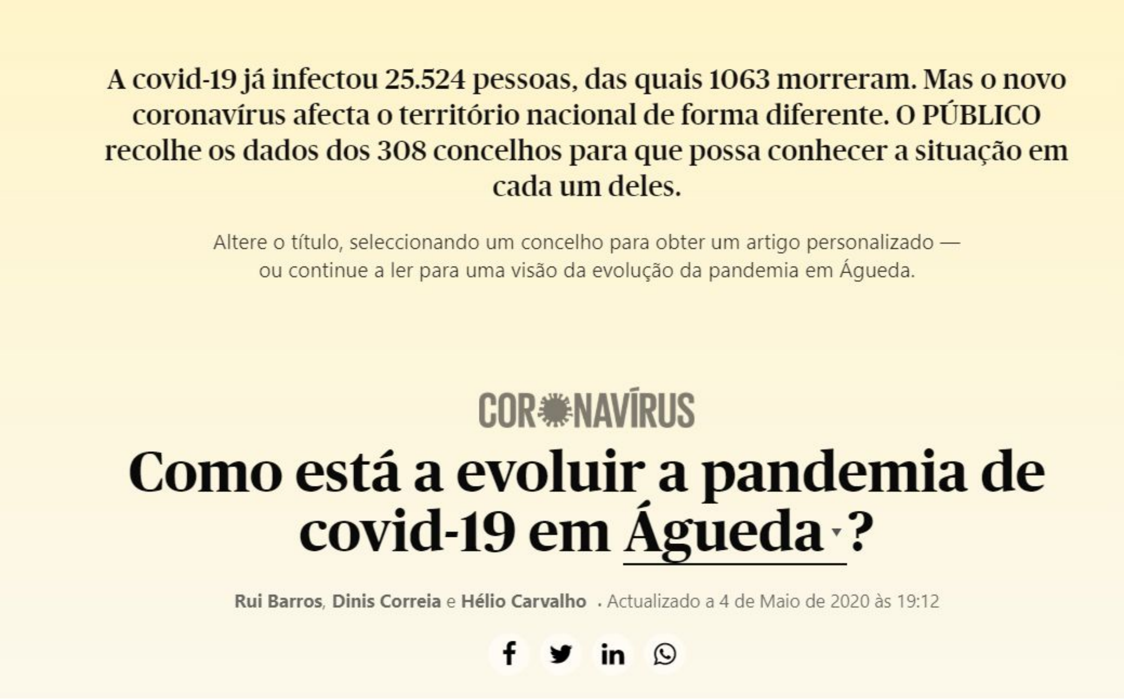


Estimates obtained by CoronaSurveys

(Updated daily)

[1] https://coronasurveys.org/

👀 Público - Personalized news per municipality

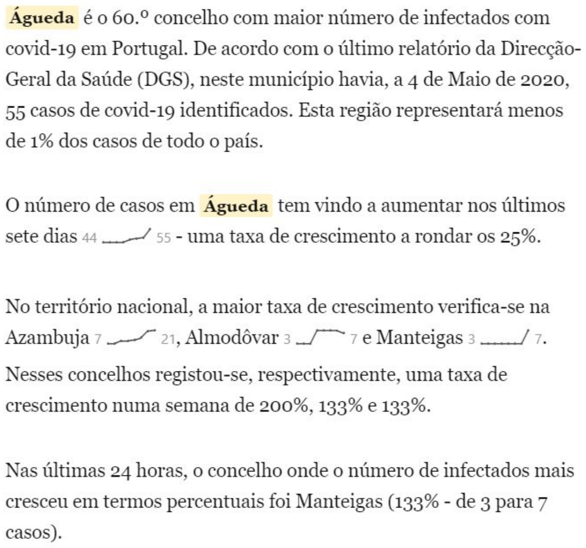


A covid-19 já infectou 25.524 pessoas, das quais 1063 morreram. Mas o novo coronavírus afecta o território nacional de forma diferente. O PÚBLICO recolhe os dados dos 308 concelhos para que possa conhecer a situação em cada um deles.

Altere o título, seleccionando um concelho para obter um artigo personalizado — ou continue a ler para uma visão da evolução da pandemia em Águeda.

Como está a evoluir a pandemia de covid-19 em Águeda ?

Rui Barros, Dinis Correia e Hélio Carvalho · Actualizado a 4 de Maio de 2020 às 19:12



Águeda é o 60.º concelho com maior número de infectados com covid-19 em Portugal. De acordo com o último relatório da Direcção-Geral da Saúde (DGS), neste município havia, a 4 de Maio de 2020, 55 casos de covid-19 identificados. Esta região representará menos de 1% dos casos de todo o país.

O número de casos em Águeda tem vindo a aumentar nos últimos sete dias 44 55 - uma taxa de crescimento a rondar os 25%.

No território nacional, a maior taxa de crescimento verifica-se na Azambuja 7 21, Almodôvar 3 7 e Manteigas 3 7. Nesses concelhos registou-se, respectivamente, uma taxa de crescimento numa semana de 200%, 133% e 133%.

Nas últimas 24 horas, o concelho onde o número de infectados mais cresceu em termos percentuais foi Manteigas (133% - de 3 para 7 casos).

(uses part of our data - total # and deaths)

[1] https://www.publico.pt/interactivo/como-esta-evoluir-pandemia-covid19-onde-vivo

👀 List of projects (partial)

- **Como achatar a curva? O que revelam as experiências dos países [1],** by Rui Barros and Dinis Correia from jornal Público
- **COVID19 Portugal by Crossroads - Portugal e um olhar sobre o mundo [2]**, by **zemanels [3]**
- **CoronaSurveys: Monitoring COVID-19 incidence via open polls [4]**, by Universidade do Minho and INESC TEC
- **COVID-19 Time varying reproduction numbers estimation for Portugal [5]**, by **Christian Perone [6]**
- + a lot more in **Project Showcase issue [7]** + **Social Networks**!

[1] https://www.publico.pt/interactivo/coronavirus-como-achatar-curva-que-revelam-experiencias-paises
[2] https://covid19.crossroads.pt/ | [3] https://github.com/zemanels
[4] https://coronasurveys.org/
[5] https://perone.github.io/covid19analysis/portugal_r0.html | [6] https://github.com/perone
[7] https://github.com/dssg-pt/covid19pt-data/issues/20

🔮 So, what’s next?

Just launched

- Number of processed testing samples, updated daily

Soon…

- Municipality data (and cartographic resources), updated daily PR #85 [1]

Interesting in helping out?

- Scrapper to automatically **detect when a new report is released**.
- Integrate mortality data from **SICO - eVM** [3].
- Participate in the **TAIKAI** challenge.
- **Your ideas**! Development is all done on GitHub, **transparently.**

[1] https://github.com/dssg-pt/covid19pt-data/pull/85 | [2] https://github.com/dssg-pt/covid19pt-data/pull/96
[3] https://evm.min-saude.pt/

🏠 Take-home message

- Society is starting to understand the power of having **transparent and open data**.

- **Automatic validation systems** could prevent mistakes like what have been seen in the last days (duplicated cases).

- To this day, our major challenge will be to guarantee our scripts continue working with the **variability of the daily reports.**

- Our offer for helping the Portuguese Health Authorities is **still standing**, and we're working on operationalizing it.

Meet us on...

dssg.pt

dssg_pt

@dssgpt

/DSSGPortugal

DATA SCIENCE FOR SOCIAL GOOD
PORTUGAL

🔴 DSSG PT !?

- DSSG PT does **not conduct groundbreaking research**.
- DSSG PT is **not a company**.
- DSSG PT is **often looking for new members**, but always for **unpaid positions**.
- DSSG PT, from a purely **technical standpoint**, is a rather **boring endeavour**.

DSSG PT !?

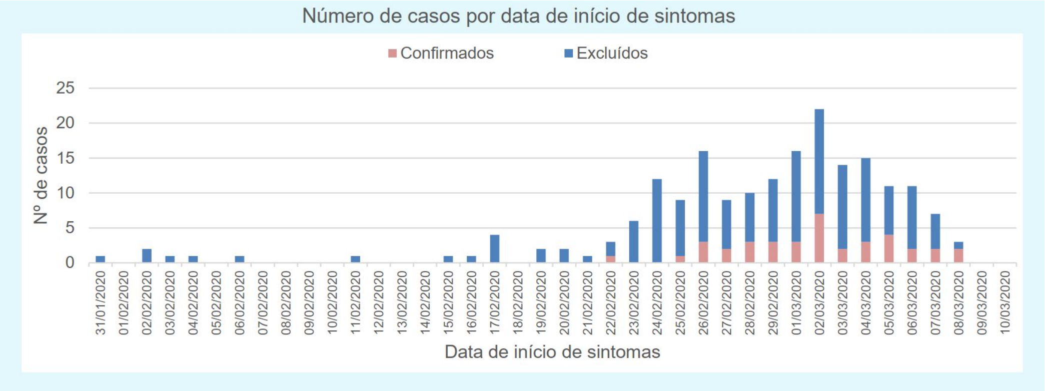
- DSSG PT always works on **real life problems**.
- DSSG PT delivers not papers, not prototypes, but usable **Data Science products**.
- DSSG PT only works on one type of problems: **those that have a positive impact for society.**

🧪 Aim: to replicate a formula

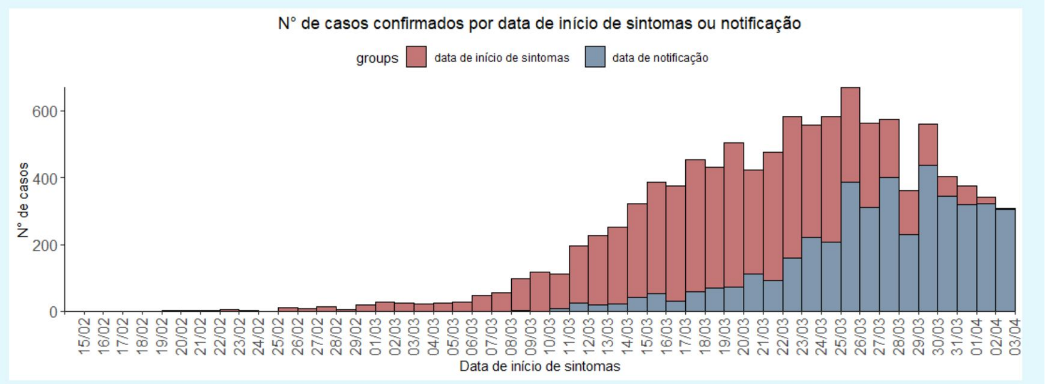
An open **community** of **data scientists**, **data lovers** and **data enthusiasts** that wants to **tackle problems that really matter**. We believe in the **power of data to transform our society** for the better, for everyone.

Beneficiaries + **Volunteers** = **Projects**

Constant changes in the reported clinical indicators



Report #8



Caso não exista informação disponível sobre data de início de sintomas, é considerada a data de diagnóstico.

Report #33