experts club

Coleta e preparação de dataset em Python e Pandas

14/10/2021

Agenda

experts club

- Sobre mim e a minha relação com o código;
- Sobre a aula e o que será entregue no final;
- Requisitos, ambiente e recursos;
- Considerações sobre o dataset para análises;
- Acesso à fonte de dados e características do site;
- Limpeza, preparação e enriquecimento do dataset;
- Limitações de sites públicos e alternativas de consulta.

Sobre mim e a minha relação com o código

experts club

- **Sergio Siqueira;**
- **Engenheiro Eletricista com ênfase em eletrônica;**
- **35 anos de experiência em TI, infraestrutura e hardware;**
- **Desenvolvimento de software como hobby e recentemente parte do trabalho;**
- **Head of devops e consultor em tecnologia;**
- **Redes sociais:**
  - https://app.rocketseat.com.br/me/sergio-siqueira-05693;
  - https://www.linkedin.com/in/snsergio/;
  - https://github.com/snsergio;

Sobre a aula e o que será entregue no final

experts club

- **Preparação de dataset em Python e Pandas;**
- **Ao final teremos um dataset otimizado e pronto para análise de dados;**

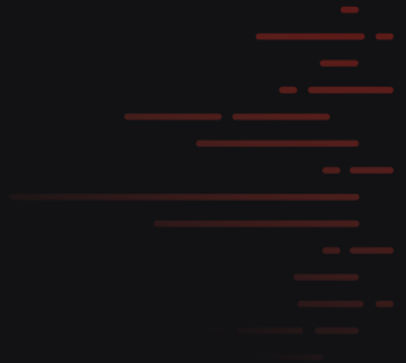
		muni	codIbge	data	semEpid	popEstim	confAcc	confAcc100k	obitoAcc	tipoLocal	uf	confDia	obitoDia
muni	semEpid												
Abadia de Goiás	202021	Abadia de Goiás	5200050.0	2020-05-23	202021	8958.0	6	66.97924	0	city	GO	5	0
	202022	Abadia de Goiás	5200050.0	2020-05-30	202022	8958.0	6	66.97924	0	city	GO	1	0
	202023	Abadia de Goiás	5200050.0	2020-06-06	202023	8958.0	8	89.30565	0	city	GO	1	0
	202024	Abadia de Goiás	5200050.0	2020-06-13	202024	8958.0	14	156.28489	0	city	GO	7	0
	202025	Abadia de Goiás	5200050.0	2020-06-20	202025	8958.0	22	245.59053	0	city	GO	8	0
	202026	Abadia de Goiás	5200050.0	2020-06-27	202026	8958.0	33	368.38580	0	city	GO	11	0
	202027	Abadia de Goiás	5200050.0	2020-07-04	202027	8958.0	42	468.85466	0	city	GO	8	0
	202028	Abadia de Goiás	5200050.0	2020-07-11	202028	8958.0	84	937.70931	0	city	GO	43	0
	202029	Abadia de Goiás	5200050.0	2020-07-18	202029	8958.0	103	1149.81023	0	city	GO	19	0
	202030	Abadia de Goiás	5200050.0	2020-07-25	202030	8958.0	135	1507.03282	1	city	GO	32	1
	202031	Abadia de Goiás	5200050.0	2020-08-01	202031	8958.0	182	2031.70351	1	city	GO	47	0
	202032	Abadia de Goiás	5200050.0	2020-08-08	202032	8958.0	217	2422.41572	1	city	GO	35	0
	202033	Abadia de Goiás	5200050.0	2020-08-15	202033	8958.0	311	3471.75709	2	city	GO	94	1
	202034	Abadia de Goiás	5200050.0	2020-08-22	202034	8958.0	357	3985.26457	2	city	GO	46	0
	202035	Abadia de Goiás	5200050.0	2020-08-29	202035	8958.0	393	4387.13999	3	city	GO	36	1
	202036	Abadia de Goiás	5200050.0	2020-09-05	202036	8958.0	442	4934.13708	4	city	GO	48	1
	202037	Abadia de Goiás	5200050.0	2020-09-12	202037	8958.0	482	5380.66533	6	city	GO	41	2
	202038	Abadia de Goiás	5200050.0	2020-09-19	202038	8958.0	508	5670.90868	7	city	GO	24	1
	202039	Abadia de Goiás	5200050.0	2020-09-26	202039	8958.0	535	5972.31525	7	city	GO	28	0
	202040	Abadia de Goiás	5200050.0	2020-10-03	202040	8958.0	579	6463.49632	7	city	GO	45	0

- https://github.com/rocketseat-experts-club/preparacao-dataset-python-pandas-2021-10-22
- Arquivos disponíveis no github:
- Jupyter notebook e script python

Requisitos, ambiente e recursos

experts club

- **Requisitos para um melhor aproveitamento da aula:**
  - Jupyter notebook;
  - Ou VScode;
  - Python 3.8 e Biblioteca Pandas.
- **Ambiente e recursos necessários:**
  - Python 3.8;
  - Pandas;
  - IDE Python.



Tópico relacionado ao conteúdo

experts club

brasil.io

City
city_ibge_code
date
epidemiological_week
estimated_population
estimated_population_2019
is_last
is_repeated
last_available_confirmed
last_available_confirmed_per_100k_inhabitants
last_available_date
last_available_death_rate
last_available_deaths
order_for_place
place_type
State
new_confirmed
new_deaths

Dataset (caso_full.csv)
epidemiological_week
date
state
city
city_ibge_code
place_type
last_available_confirmed
last_available_confirmed_per_100k_inhabitants
estimated_population
new_confirmed
last_available_deaths
new_deaths

OpenCage

Nossa tabela
Semana epidemiológica
Data
UF
Município
Código IBGE
Tipo do local
Casos confirmados acumulado
Casos confirmados acumulado por 100k habitantes
População estimada
Casos confirmados no dia
Óbitos acumulados
Óbitos no dia
Latitude
Longitude

experts club

Obrigado!

Sergio Siqueira
sergio@tecnosiq.dev
https://app.rocketseat.com.br/me/sergio-siqueira-05693
https://www.linkedin.com/in/snsergio
https://github.com/snsergio

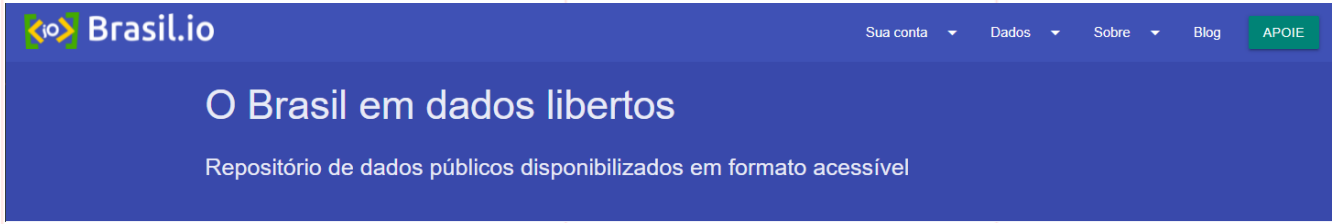
Coleta de dados para análises em Python

experts club

- **Considerações sobre o dataset para análises**
  - Entendimento da construção do dataset e frequência de coleta
  - Definição das informações necessárias para a análise a ser feita
- **Acesso à fonte de dados e características de acesso**
- **Limpeza e preparação do dataset e as razões para a otimização**
- **Enriquecimento do dataset para resultados mais abrangentes**
- **Alternativas à limitações em sites com informações públicas**
  - Salvar arquivos com informações que não mudam, por exemplo

brasil.io

experts club



COVID-19

Boletins informativos e casos do coronavírus por município por dia

Fonte original: Secretarias de Saúde estaduais
Libertado por: Álvaro Justen e dezenas de colaboradores
Código-fonte: https://github.com/turicas/covid19-br
Licença: Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Links relacionados: Boletins PR, Boletins SP, Boletins RO, Boletins MG, Boletins RS, Boletins MT, Boletins MS, Boletins BA, Boletins PE, Informações sobre a coleta de dados (manual), Boletins AC, Boletins AL, Boletins AM, Boletins AP, Boletins CE, Boletim ES, Boletins GO, Boletins MA, Boletins PA, Boletins PB, Boletins PI, Boletins RJ, Boletins RN, Boletins RR, Boletins SC (1), Boletins SC (2), Boletins DF (1), Boletins DF (2), Boletins SE, Boletins TO, Boletins RJ (2), Documentação da API, Perguntas e respostas sobre os dados, Portal da Transparência do Registro Civil

Tabelas: boletim, caso, **caso_full**, obito_cartorio.

Informações úteis

Essa tabela possui os casos confirmados e óbitos obtidos dos boletins das **Secretarias Estaduais de Saúde** (SES). Os dados foram enriquecidos, de forma que a partir do momento em que um município confirma um caso, ele sempre aparecerá nessa tabela (mesmo que para uma determinada data a SES não tenha liberado o boletim - nesse caso é repetido o dado do dia anterior). Caso queira ver a tabela original (sem repetição e com datas faltantes), visite caso.
Além de acessar os dados por essa interface você pode também baixar o dataset completo ou acessá-lo via API.
Acesse também o **Painel COVID-19**.
ATENÇÃO: antes de nos enviar dúvidas sobre os dados:

Saiba mais sobre esse dataset e o projeto Brasil.IO COVID-19;
Leia a documentação dessa tabela;
Leia nossa FAQ sobre esse dataset;
Caso tenha verificado dados incorretos, verifique se nas "Observações" do boletim para o dia em questão existem informações sobre;
Caso você não esteja conseguindo abrir o CSV no Excel, aprenda a fazê-lo.

Como acessar os dados

experts club

- **Consulta via API**

Dataset Data List

GET

Importante: Use a API pública com parcimônia para não onerar nossos servidores.

Estamos tendo um volume muito grande de chamadas a API para interagir com datasets completos. Caso você precise manipular volumes muito grandes de dados, baixe o CSV usando os links disponíveis **nos datasets do Brasil.IO**. Esses links usam nosso servidor de arquivos estáticos e são a maneira mais rápida (e correta) de baixar bases de dados completas.

Em breve limitaremos o acesso automatizado ao site, dados esses abusos que vem acontecendo. Peço que acompanhe as novidades em nosso canal no Telegram.

```
GET /dataset/covid19/caso_full/data/
```

- **Arquivo CSV**

dataset: dataset covid19

Data de captura: 2021-09-13

Colabore!

O Brasil.IO tem como objetivo facilitar o acesso a dados públicos brasileiros. O projeto é desenvolvido de forma colaborativa, todo o código está disponível como software livre e os custos são pagos através de uma campanha de financiamento coletivo. Se você acredita nos ideais do projeto ou esses dados estão sendo úteis para você ou sua empresa, considere fazer uma doação ou colaborar de outras formas.

Baixe os arquivos:

Arquivo	Tamanho	SHA512SUM
boletim.csv.gz	186.09kB	687ba4e2a3376f2f1b0a26fc32aa8a8ff3715ed3249d34af6bc1f86af5b6d853d97301d8fd906711834fc72cce03d58d9715fff77c0061a24f649aac9925b14a
caso.csv.gz	26.92MB	7e1ff83d210c5b0bcd6d31e64fc5d38816b15620e78a6b724a86c126721108a88a4e5b1df31dc17b080d6bfdd477352655c235ffe78dadb2c95af13c114a596a
caso_full.csv.gz	64.15MB	3167615e5e2a42d5558a2e4f0660e2ed50415b86d6acaae25a7e79b399bf07a9e4431e5bce72f843e1c26cf917d5366122bc26a9f6444e6da1d0b5d4f1ddb9a4
obito_cartorio.csv.gz	434.18kB	94f91c9e05516b8f72419a4937d3d4aa27989d64e76b9badabec0b496e6ac8b96c6dccf4ab116f79d175eae93ac93fb1436006013f515034feb8600dd8c940e0
SHA512SUMS	586B	c5ee100e5d7c1c3b6dbf1f8fcb14bd101ddd22c69b4925bf8204281f97d2af7b869ae9df44f6265cf1f5362230ae77e0683d48736497c455e6b148b8de6a1f13

Campos do dataset

experts club

city: nome do município (pode estar em branco quando o registro é referente ao estado, pode ser preenchido com Importados/Indefinidos também).

city_ibge_code: código IBGE do local.

date: data de coleta dos dados no formato YYYY-MM-DD.

epidemiological_week: número da semana epidemiológica no formato YYYYWW.

estimated_population: população estimada para esse município/estado em 2020, segundo o IBGE. (acesse o script que faz o download e conversão dos dados de população).

estimated_population_2019: população estimada para esse município/estado em 2019, segundo o IBGE. ATENÇÃO: essa coluna possui valores desatualizados, prefira usar a coluna estimated_population.

is_last: campo pré-computado que diz se esse registro é o mais novo para esse local, pode ser True ou False (caso filtre por esse campo, use is_last=True ou is_last=False, não use o valor em minúsculas).

is_repeated: campo pré-computado que diz se as informações nesse registro foram publicadas pela Secretaria Estadual de Saúde no dia date ou se o dado é repetido do último dia em que o dado está disponível (igual ou anterior a date). Isso ocorre pois nem todas as secretarias publicam boletins todos os dias. Veja também o campo last_available_date.

last_available_confirmed: número de casos confirmados do último dia disponível igual ou anterior à data date.

last_available_confirmed_per_100k_inhabitants: número de casos confirmados por 100.000 habitantes (baseado em estimated_population) do último dia disponível igual ou anterior à data date.

last_available_date: data da qual o dado se refere.

last_available_death_rate: taxa de mortalidade (mortes / confirmados) do último dia disponível igual ou anterior à data date.

last_available_deaths: número de mortes do último dia disponível igual ou anterior à data date.

order_for_place: número que identifica a ordem do registro para este local. O registro referente ao primeiro boletim em que esse local aparecer será contabilizado como 1 e os demais boletins incrementarão esse valor.

place_type: tipo de local que esse registro descreve, pode ser city ou state.

state: sigla da unidade federativa, exemplo: SP.

new_confirmed: número de novos casos confirmados desde o último dia (note que caso is_repeated seja True, esse valor sempre será 0 e que esse valor pode ser negativo caso a SES remaneje os casos desse município para outro).

new_deaths: número de novos óbitos desde o último dia (note que caso is_repeated seja True, esse valor sempre será 0 e que esse valor pode ser negativo caso a SES remaneje os casos desse município para outro).

Campos utilizados do dataset

experts club

city: nome do município (pode estar em branco quando o registro é referente ao estado, pode ser preenchido com Importados/Indefinidos também).

city_ibge_code: código IBGE do local.

date: data de coleta dos dados no formato YYYY-MM-DD.

epidemiological_week: número da semana epidemiológica no formato YYYYWW.

estimated_population: população estimada para esse município/estado em 2020, segundo o IBGE. (acesse o script que faz o download e conversão dos dados de população).

estimated_population_2019: população estimada para esse município/estado em 2019, segundo o IBGE. ATENÇÃO: essa coluna possui valores desatualizados, prefira usar a coluna estimated_population.

is_last: campo pré-computado que diz se esse registro é o mais novo para esse local, pode ser True ou False (caso filtre por esse campo, use is_last=True ou is_last=False, não use o valor em minúsculas).

is_repeated: campo pré-computado que diz se as informações nesse registro foram publicadas pela Secretaria Estadual de Saúde no dia date ou se o dado é repetido do último dia em que o dado está disponível (igual ou anterior a date). Isso ocorre pois nem todas as secretarias publicam boletins todos os dias. Veja também o campo last_available_date.

last_available_confirmed: número de casos confirmados do último dia disponível igual ou anterior à data date.

last_available_confirmed_per_100k_inhabitants: número de casos confirmados por 100.000 habitantes (baseado em estimated_population) do último dia disponível igual ou anterior à data date.

last_available_date: data da qual o dado se refere.

last_available_death_rate: taxa de mortalidade (mortes / confirmados) do último dia disponível igual ou anterior à data date.

last_available_deaths: número de mortes do último dia disponível igual ou anterior à data date.

order_for_place: número que identifica a ordem do registro para este local. O registro referente ao primeiro boletim em que esse local aparecer será contabilizado como 1 e os demais boletins incrementarão esse valor.

place_type: tipo de local que esse registro descreve, pode ser city ou state.

state: sigla da unidade federativa, exemplo: SP.

new_confirmed: número de novos casos confirmados desde o último dia (note que caso is_repeated seja True, esse valor sempre será 0 e que esse valor pode ser negativo caso a SES remaneje os casos desse município para outro).

new_deaths: número de novos óbitos desde o último dia (note que caso is_repeated seja True, esse valor sempre será 0 e que esse valor pode ser negativo caso a SES remaneje os casos desse município para outro).

Campos utilizados do dataset

experts club

city
city_ibge_code
date
epidemiological_week
estimated_population
estimated_population_2019
is_last
is_repeated
last_available_confirmed
last_available_confirmed_per_100k_inhabitants
last_available_date
last_available_death_rate
last_available_deaths
order_for_place
place_type
State
new_confirmed
new_deaths

Como queremos a análise?

experts club

Nossa tabela	Dataset (caso_full.csv)
Semana epidemiológica	epidemiological_week
Data	date
UF	state
Município	city
Código IBGE	city_ibge_code
Tipo do local	place_type
Casos confirmados acumulado	last_available_confirmed
Casos confirmados acumulado por 100k habitantes	last_available_confirmed_per_100k_inhabitants
População estimada	estimated_population
Casos confirmados no dia	new_confirmed
Óbitos acumulados	last_available_deaths
Óbitos no dia	new_deaths

Como queremos a análise?

experts club

Nossa tabela	Dataset (caso_full.csv)
Semana epidemiológica	epidemiological_week
Data	Date
UF	State
Município	City
Código IBGE	city_ibge_code
Tipo do local	place_type
Casos confirmados acumulado	last_available_confirmed
Casos confirmados acumulado por 100k habitants	last_available_confirmed_per_100k_inhabitants
População estimada	estimated_population
Casos confirmados no dia	new_confirmed
Óbitos acumulados	last_available_deaths
Óbitos no dia	new_deaths
Latitude	??????????
Longitude	??????????