Análise dos dados de vacinação da COVID-19 e seus estabelecimentos de saúde nas regiões norte e sudeste do Brasil

Alessandra Pierro Saraiva, Fabiane Monteiro Carvalho e Massao Oliveira Iwanaga

Universidade Federal do Rio de Janeiro
Av. Pedro Calmon no 50, 2o andar, Ilha do Fundão - Rio de Janeiro, RJ, 21941-596, Brasil
alessandrapsaraiva@gmail.com, famcrj@gmail.com,
massaoiwanaga@ufrj.br

Resumo. A humanidade vem enfrentando uma pandemia sem precedentes através da doença contagiosa chamada COVID19 (Coronavirus disease 2019). No contexto brasileiro, fazem-se necessárias análises sobre a vacinação da COVID19 para avaliar o perfil dos vacinados e de seus estabelecimentos de saúde nas regiões norte e sudeste do país devido às suas desigualdades. Serão aplicadas técnicas de Data Science (DS) nos datasets de vacinação e de estabelecimentos de saúde ativos. Este trabalho será implementado através do Google Colaboratory, um serviço de nuvem gratuito hospedado pela Google, sem nenhuma configuração local necessária e com acesso gratuito a GPUs (Unidades de Processamento Gráfico). A proveniência deste experimento permitirá a reprodutibilidade do mesmo. Ao final, serão fornecidas as visualizações das informações necessárias para tomadas de decisão.

Palavras-chave: Data Science, Analytics, Covid19 Vaccination, Provenance.

1 Introdução

A humanidade vem enfrentando uma pandemia sem precedentes através do surto da doença contagiosa chamada COVID-19 (*Coronavirus disease* 2019). No contexto brasileiro, duas regiões foram particularmente muito afetadas. Em primeiro lugar, a região norte[1], com população de aproximadamente 18,6 milhões de habitantes[2], e, em

[1] https://revistagalileu.globo.com/Ciencia/Saude/noticia/2020/06/pesquisa-aponta-regiao-norte-como-epicentro-da-covid-19-no-pais.html. Acesso em 07/05/2021.

[2] https://pt.wikipedia.org/wiki/Regi%C3%A3o_Norte_do_Brasil. Acesso em: 07/05/2021.

segundo, a região sudeste[3], a região mais populosa, com número de habitantes em torno de 87,7 milhões[4].

Com o início e avanço, no ano de 2021, da vacinação contra essa enfermidade em todo o país, fazem-se necessárias análises a partir dos dados dos vacinados e dos estabelecimentos de saúde para avaliar diversas questões como: perfis dos vacinados e dos estabelecimentos de saúde, se os sistemas de vacinações locais suprem a carência de sua população local ou se ela precisa se vacinar em outro município, discrepâncias e/ou desigualdades entre as regiões, velocidade de vacinação, entre outras.

Este trabalho, através de três *datasets* disponibilizados publicamente pelo governo brasileiro, busca analisar comparativamente os dados das regiões norte e sudeste do Brasil. Em termos de tecnologia, foram usados para chegar em tal fim: a linguagem de programação Python, o ambiente de desenvolvimento Google Colab, três *datasets* no formato CSV (Comma-Separated Values), o versionador de arquivos GIT e o serviço de hospedagem em nuvem Google Drive. Foram implementados também os conceitos de proveniência (mediante a biblioteca PROV) e reprodutibilidade, como descrito ulteriormente.

Verifica-se que essas análises poderão contribuir para um estudo mais detalhado sobre a cobertura da vacinação nessas regiões com seus quantitativos e perfis, permitindo tomadas de decisões através dos dados apresentados para aprimorar o processo de vacinação nesses locais, com informações sobre os estabelecimentos de saúde, de forma a diminuir a aglomeração de pessoas no atendimento e também comparar a distribuição desses estabelecimentos nas duas regiões do país.

O trabalho está organizado em seis seções, já considerando esta introdução. Na segunda seção, serão abordadas sucintamente as referências bibliográficas consultadas para o referencial teórico. Na terceira, serão apresentados os *datasets* usados na implementação. Na quarta seção, será descrito o estudo de caso em si, que, através da manipulação dos *datasets*, gerou os gráficos, as análises e também as informações de proveniência. Na quinta, apresenta-se uma discussão acerca dos resultados obtidos na seção anterior. Para finalizar, o trabalho se encerra com a conclusão, onde é feito o fechamento das ideias discutidas.

1.1 **Objetivos gerais**

Este trabalho almeja aplicar técnicas de Data Science (DS) que subsidiem análises do problema, tais como:

- ❏ Importação, limpeza e tratamento de dados dos *datasets*;
- ❏ Análise exploratória dos *datasets*;
- ❏ Extração do conhecimento dos *datasets*; e
- ❏ Visualização de dados que auxiliem na análise do problema proposto.

Além do listado anteriormente, a implementação deverá prever a reutilização do código, bem como informações de proveniência sobre os dados utilizados.

---

[3] http://associacaopaulistamedicina.org.br/noticia/covid-19-sudeste-e-a-regiao-mais-afetada-do-pais. Acesso em 07/05/2021.

[4] https://pt.wikipedia.org/wiki/Regi%C3%A3o_Sudeste_do_Brasil. Acesso em: 07/05/2021.

1.2 **Objetivos específicos**

Quanto aos objetivos específicos, o presente artigo busca fornecer informações sobre os vacinados e também dos seus locais de vacinação por município, estados, regiões metropolitanas e interioranas das regiões norte e sudeste, da seguinte forma:

1. Perfil dos vacinados (sexo, idade, raça, grupo) e quantitativos por estado nas duas regiões, dividido entre regiões metropolitanas e regiões interioranas;
2. Perfil dos vacinados (sexo, idade, raça, grupo) e quantitativos por tipo de estabelecimento de saúde por estado nas duas regiões;
3. Quantitativo dos estabelecimentos de saúde e seus tipos nos estados nas duas regiões por total de população;
4. Percentual de vacinados que tomaram a vacina em município diferente do da sua residência mas no mesmo estado; e
5. Visualização de % da população vacinada ao longo do tempo, nas duas regiões, permitindo uma análise comparativa da evolução da vacinação entre as regiões. (remover?)

2 Trabalhos Relacionados

Em [5], Marcelo Ribeiro contribui analiticamente com o debate acerca das desigualdades de renda nas regiões metropolitanas brasileiras que, normalmente, tendem a considerar somente as características dos indivíduos como cor/raça e sexo. Marcelo defende que o grau de escolaridade tem uma correlação importante e deve ser uma variável a ser analisada no contexto de desigualdade.

Já Luiza Garnelo e Sully Sampaio, em [6], discutem as bases sócio-culturais do controle social em saúde indígena, considerando o risco negativo de se terceirizar a saúde indígena para o mercado privado. O trabalho é importante devido a quantidade proporcionalmente grande de indígenas vivendo na região norte do Brasil. Os indígenas são etnias que tiveram prioridade na vacinação contra a COVID-19.

O artigo de Maria Donalisio et. al [7] disserta, no contexto da vacinação contra influenza em idosos em municípios do sudeste, sobre os fatores associados à essa vacinação. Ela objetiva defender que, mesmo gratuita e disponível no Brasil desde 1999, a cobertura vacinal contra a influenza é inadequada em diversos municípios do país. Através dessa exposição, buscou estimular a cobertura vacinal e identificar fatores relacionados à vacinação contra a influenza em idosos. Apesar da disparidade de gravidade entre a influenza e a COVID-19, todo o aprendizado obtido através do estudo do surto da influenza e de como foi e é gerida sua vacinação, pode ser útil para ser aprimorado no combate a pandemia na qual vivemos.

(se puderem escrever sobre mais uma referência e concatenar as ideias para entre os parágrafos)

- Território e Desigualdades de Renda em Regiões Metropolitanas do Brasil

    - https://www.scielo.br/pdf/dados/v58n4/0011-5258-dados-58-4-0913.pdf
- Acesso e cobertura da Atenção Primária à Saúde para populações rurais e urbanas na região norte do Brasil
    - https://www.scielosp.org/pdf/sdeb/2018.v42nspe1/81-99/pt
- Bases sócio-culturais do controle social em saúde indígena. Problemas e questões na Região Norte do Brasil
    - https://www.scielosp.org/pdf/csp/2003.v19n1/311-317/pt
- Avaliação da assistência à saúde da mulher e da criança em localidade urbana da região Sudeste do Brasil
    - https://www.scielosp.org/article/rsp/2000.v34n3/266-271/pt/
- Atenção primária à saúde e Organizações Sociais nas capitais da Região Sudeste do Brasil: 2009 e 2014
    - https://www.scielosp.org/article/csp/2019.v35n4/e00089118/
- Mortes evitáveis por ações do Sistema Único de Saúde na população da Região Sudeste do Brasil
    - https://www.scielosp.org/article/csc/2019.v24n3/887-898/pt/
- Autopercepção de saúde e aspectos clínico-funcionais dos idosos atendidos em uma unidade básica de saúde no norte do Brasil
    - https://periodicos.unichristus.edu.br/jhbs/article/viewFile/1054/412
- Fatores associados à vacinação contra influenza em idosos em município do Sudeste do Brasil
    - https://www.scielosp.org/pdf/rsp/2006.v40n1/115-119/pt
    -

3 Datasets

Neste trabalho, serão usados três *datasets* conforme descritos em seguida. Eles foram obtidos através de duas fontes: o Portal Brasileiro de Dados Abertos[5], ferramenta disponibilizada pelo governo para que todos os brasileiros possam encontrar e utilizar os dados abertos nacionais e informações públicas; e o site do Instituto Brasileiro de Geografia e Estatística (IBGE).

3.1 Dataset da Campanha Nacional de Vacinação contra COVID-19 [2]

O primeiro *dataset* utilizado é o maior e possui informações sobre os vacinados (idade, data de nascimento, sexo, raça e etc.) contra a COVID-19, os estabelecimentos de saúde (nome, município, estado e etc.) onde ocorreu a vacinação, grupo de atendimento ao qual o vacinado pertence e da vacina (fabricante, lote e etc.), além de outras. Ele possui relacionado um dicionário de dados, conforme exibido a seguir, contendo os 32 campos, suas descrições e categorias.

[5] https://dados.gov.br/

Ordem	Campo	Descrição	Categoria
1	document_id	Identificador do documento	
2	paciente_id	Identificador do vacinado	
3	paciente_idade	Idade do vacinado	
4	paciente_dataNascimento	Data de nascimento do vacinado	
5	paciente_enumSexoBiologico	Sexo do vacinado	M = Masculino, F = Feminino
6	paciente_racaCor_codigo	Código da raça/cor do vacinado	1; 2; 3; 4; 99
7	paciente_racaCor_valor	Descrição da raça/cor do vacinado	1 = Branca; 2 = Preta; 3 = Parda; 4 = Amarela; 99 = Sem informação
8	paciente_endereco_coIbgeMunicipio	Código IBGE do município de endereço do vacinado	
9	paciente_endereco_coPais	Código do país de endereço do vacinado	
10	paciente_endereco_nmMunicipio	Nome do município de endereço do vacinado	
11	paciente_endereco_nmPais	Nome do país de endereço do vacinado	
12	paciente_endereco_uf	Sigla da UF de endereço do vacinado	
13	paciente_endereco_cep	5 dígitos para anonimizado e 7 dígitos para identificado	
14	paciente_nacionalidade_enumNacionalidade	Nacionalidade do vacinado	
15	estabelecimento_valor	Código do CNES do estabelecimento que realizou a vacinação	
16	estabelecimento_razaoSocial	Nome/Razão Social do estabelecimento	
17	estabelecimento_noFantasia	Nome fantasia do estabelecimento	
18	estabelecimento_municipio_codigo	Código do município do estabelecimento	
19	estabelecimento_municipio_nome	Nome do município do estabelecimento	
20	estabelecimento_uf	Sigla da UF do estabelecimento	
21	vacina_grupo_atendimento_code	Código do grupo de atendimento ao qual pertence o vacinado	
22	vacina_grupo_atendimento_nome	Nome do grupo de atendimento ao qual pertence o vacinado	
23	vacina_categoria_code	Código da categoria	
24	vacina_categoria_nome	Descrição da categoria	
25	vacina_lote	Número do lote da vacina	
26	vacina_fabricante_nome	Nome do fabricante/fornecedor	
27	vacina_fabricante_referencia	CNPJ do fabricante/fornecedor	
28	vacina_dataAplicacao	Data de aplicação da vacina	
29	vacina_descricao_dose	Descrição da dose	
30	vacina_codigo	Código da vacina	
31	vacina_nome	Nome da vacina/produto	
32	sistema origem	Nome do sistema de origem	

Fig. 1. Dicionário de dados do *dataset* da Campanha Nacional de Vacinação contra COVID-19.

É importante notar, que como a vacinação contra o COVID-19 continua em andamento, o *dataset* está sendo frequentemente atualizado no portal, cada vez com mais dados. Decidimos fazer o trabalho em cima da versão do *dataset* disponibilizada no dia 05/03/2021. Os dados em si, são em um total de 7.958.949 registros, incluído o cabeçalho, e 4.2 GB de tamanho físico, disponibilizados em formato CSV.

3.2 **Cadastro Nacional de Estabelecimentos de Saúde (CNES) - Ativo [1]**

O segundo *dataset* possui informações sobre todos os estabelecimentos de saúde (código, nome, endereço, tipo e etc.) ativos no país, sejam públicos ou privados. Apesar de não haver disponível um dicionário de dados oficial para ele, criamos um, conforme exibido a seguir, a partir de informações contidas na página do *dataset* e no cabeçalho do arquivo baixado.

Ordem	Campo	Descrição
1	co_cnes	Código CNES
2	co_ibge	Código IBGE
3	no_fantasia	Nome Fantasia
4	ds_tipo_unidade	Tipo de estabelecimento de saúde: Posto de Saúde, Centro de Saúde/Unidade Básica de Saúde, Policlínica, Hospital Geral,Hospital Especializado, Unidade Mista, Pronto Socorro Geral, Pronto Socorro Especializado, Consultório Isolado, UnidadeMóvel Fluvial, Clínica Especializada/Amb. Especializado, Unidade de Serviço de Apoio de Diagnose e Terapia, Unidade MóvelTerrestre, Unidade Móvel de Nível Pré-hospitalar na Área de Urgência e Emergência: Farmácia: Unidade de Vigilância em Saúde,Cooperativa, Centro de Parto Normal Isolado,Hospital /Dia-Isolado, Central de Regulação de Serviços de Saúde, LaboratórioCentral de Saúde Pública – LACEN, Secretaria de Saúde
5	tp_gestao	Tipo de gestão: Estadual (E), Municipal (M) ou Dupla (D)
6	no_logradouro	Nome do Logradouro
7	nu_endereco	Número do Endereço
8	no_bairro	Bairro
9	co_cep	CEP
10	uf	UF do estabelecimento
11	municipio	Município
12	nu_telefone	Número de Telefone

Fig. 2. Dicionário de dados do *dataset* do Cadastro Nacional de Estabelecimentos de Saúde (CNES) - Ativo.

Os dados em si, são em um total de 354.805 registros, incluído o cabeçalho, e 56 MB de tamanho físico, disponibilizados em formato CSV.

3.3 Regiões Metropolitanas das Regiões Sudeste e Norte do Brasil

O terceiro e último *dataset* foi gerado manualmente a partir do *dataset* de regiões metropolitanas do Brasil disponibilizado pelo Instituto Brasileiro de Geografia e Pesquisa [3]. Foram filtrados os municípios componentes dos estados das regiões sudeste e norte do Brasil que fossem considerados de região metropolitana. O dicionário de dados apresentado a seguir, também criado, contém a lista dos campos do arquivo e suas descrições.

Ordem	Campo	Descrição
1	REGIAO	Região do Brasil (NORTE ou SUDESTE)
2	UF	Código da Unidade Federativa
3	MUNICIPIO	Nome do Município
4	COD_MUNICIPIO	Código IBGE do Município

Fig. 3. Dicionário de dados do *dataset* de Regiões Metropolitanas das Regiões Sudeste e Norte do Brasil.

É importante frisar que o estado do Acre (AC) oficialmente não possui região metropolitana[6], sendo assim, no escopo deste trabalho, para não haver a desconsideração de um estado, foi decidido que o município de Rio Branco, a capital, fosse considerado como região metropolitana do Acre. Essa linha foi incluída no *dataset*.

4 Estudo de Caso

A implementação deste trabalho ocorreu através do Google Colab. O Google Colaboratory, ou simplesmente, Colab, é um serviço de nuvem gratuito hospedado pela Google para, dentre outros fins, escrever código Python no navegador sem nenhuma configuração local necessária e com acesso gratuito a GPUs (Unidades de Processamento Gráfico)[7]. Os *datasets* foram baixados dos repositórios conforme indicado no capítulo 3 e em seguida colocados no Google Drive, de onde foram acessados pelo *notebook* do Colab.

Algumas bibliotecas foram usadas no projeto, como por exemplo a Pandas, que é uma biblioteca de software escrita para a linguagem de programação Python com fins de manipulação e análise de dados. Em particular, ele oferece estruturas de dados e operações para manipular tabelas numéricas e séries temporais[8]. Ela é a responsável por importar os arquivos, uma vez que ambos *datasets* estão no formato CSV. Outra biblioteca usada é a NumPy, útil para realização de cálculos numéricos em *arrays* multidimensionais. E, por último, vale ressaltar a utilização da biblioteca matplotlib.pyplot para a confecção dos gráficos analíticos.

Alguns *dataframes* foram criados para possibilitar as análises. Primeiramente, foi feito o *merge* entre os dados de vacinação e os dados de CNES através do *join* pelo campo “Código CNES”. A partir de então, foram gerados os *dataframes* das regiões norte e sudeste através do filtro de suas UFs correspondentes. Derivados dos dados do norte e sudeste, foram gerados os *dataframes* de suas respectivas regiões metropolitanas e interioranas. Para isso, tomou-se como premissa que o interior de um estado é toda a região não pertencente à região metropolitana.

4.1 **Análise e seleção dos dados**

Os sete estados da região norte e os quatro da região sudeste, assim como suas respectivas siglas, estão listados a seguir nas Tabelas 1 e 2.

---

[6] http://prvl.org.br/regioes-metropolitanas/regioes-metropolitanas/ Acesso em: 07/05/2021.

[7] https://colab.research.google.com/

[8] https://en.wikipedia.org/wiki/Pandas_(software)

Tabela 1. Estados da região norte do Brasil.

Amazonas (AM)	Acre (AC)	Amapá (AP)
Pará (PA)	Rondônia (RO)	Roraima (RR)
Tocantins (TO)		

Tabela 2. Estados da região sudeste do Brasil.

Espírito Santo (ES)	Minas Gerais (MG)
Rio de Janeiro (RJ)	São Paulo (SP)

Como os *datasets* de vacinação e de CNES possuem informações sobre todos os estados brasileiros, com o intuito de reduzir a quantidade de dados a serem analisados e trazer eventuais ganhos de performance, foi feita uma seleção inicial nos *dataframes* para utilizar apenas das informações dos estados das regiões norte e sudeste, escopo deste trabalho. Ou seja, no *dataset* de vacinação, apenas foram considerados os registros cujo campo “estabelecimento_uf” contivessem um dos seguintes valores: 'RJ', 'SP','MG', 'ES', 'AM', 'AC', 'AP', 'PA', 'RO', 'RR', 'TO'.

Os campos considerados irrelevantes para atingir os objetivos desta pesquisa foram eliminados dos *datasets*. Por exemplo: No *dataset* dos vacinados, optou-se por excluir os seguintes campos: “CEP do Paciente”, “Lote da Vacina”, “Código da Vacina”, “Razão Social do Estabelecimento”, “Nome Fantasia do Estabelecimento”, “Referência do Fabricante da Vacina” e “Sistema Origem”, conforme destacado na Figura 3 a seguir.

Ordem	Descrição	Categoria	Categoria
1	document_id	Identificador do documento	
2	paciente_id	Identificador do vacinado	
3	paciente_idade	Idade do vacinado	
4	paciente_dataNascimento	Data de nascimento do vacinado	
5	paciente_enumSexoBiologico	Sexo do vacinado	M: masculino; F: feminino
6	paciente_racaCor_codigo	Código da raça/cor do vacinado	1; 2; 3; 4; 99
7	paciente_racaCor_valor	Descrição da raça/cor do vacinado	1: branca; 2: preta; 3: parda; 4: amarela; 99: sem informação
8	paciente_endereco_coIbgeMunicipio	Código IBGE do município de endereço do vacinado	
9	paciente_endereco_coPais	Código do país de endereço do vacinado	
10	paciente_endereco_nmMunicipio	Nome do município de endereço do vacinado	
11	paciente_endereco_nmPais	Nome do país de endereço do vacinado	
12	paciente_endereco_uf	Sigla da UF de endereço do vacinado	
13	paciente_endereco_cep	5 dígitos para anonimizado e 7 dígitos para identificado	
14	paciente_nacionalidade_enumNacionalidade	Nacionalidade do vacinado	
15	estabelecimento_valor	Código do CNES do estabelecimento que realizou a vacinação	
16	estabelecimento_razaoSocial	Nome/Razão Social do estabelecimento	
17	estalecimento_noFantasia	Nome fantasia do estabelecimento	
18	estabelecimento_municipio_codigo	Código do município do estabelecimento	
19	estabelecimento_municipio_nome	Nome do município do estabelecimento	
20	estabelecimento_uf	Sigla da UF do estabelecimento	
21	vacina_grupoAtendimento_codigo	Código do grupo de atendimento ao qual pertence o vacinado	
22	vacina_grupoAtendimento_nome	Nome do grupo de atendimento ao qual pertence o vacinado	
23	vacina_categoria_codigo	Código da categoria	
24	vacina_categoria_nome	Descrição da categoria	
25	vacina_lote	Número do lote da vacina	
26	vacina_fabricante_nome	Nome do fabricante/fornecedor	
27	vacina_fabricante_referencia	CNPJ do fabricante/fornecedor	
28	vacina_dataAplicacao	Data de aplicação da vacina	
29	vacina_descricao_dose	Descrição da dose	
30	vacina_codigo	Código da vacina	
31	vacina_nome	Nome da vacina/produto	
32	sistema_origem	Nome do sistema de origem	

Fig. 3. Colunas (em vermelho) excluídas do *dataset* de Vacinação.

No *dataset* de CNES, os campos eliminados foram: “Nome do Logradouro”, “Código IBGE”, “Número do Endereço”, “Bairro” e “CEP”, conforme destacado na Figura 4 a seguir.

Ordem	Campo	Descrição
1	co_cnes	Código CNES
2	co_ibge	Código IBGE
3	no_fantasia	Nome Fantasia
4	ds_tipo_unidade	Tipo de estabelecimento de saúde: Posto de Saúde, Centro de Saúde/Unidade Básica de Saúde, Policlínica, Hospital Geral,Hospital Especializado, Unidade Mista, Pronto Socorro Geral, Pronto Socorro Especializado, Consultório Isolado, UnidadeMóvel Fluvial, Clínica Especializada/Amb. Especializado, Unidade de Serviço de Apoio de Diagnose e Terapia, Unidade MóvelTerrestre, Unidade Móvel de Nível Pré-hospitalar na Área de Urgência e Emergência: Farmácia: Unidade de Vigilância em Saúde,Cooperativa, Centro de Parto Normal Isolado,Hospital /Dia-Isolado, Central de Regulação de Serviços de Saúde, LaboratórioCentral de Saúde Pública – LACEN, Secretaria de Saúde
5	tp_gestao	Tipo de gestão: Estadual (E), Municipal (M) ou Dupla (D)
6	no_logradouro	Nome do Logradouro
7	nu_endereco	Número do Endereço
8	no_bairro	Bairro
9	co_cep	CEP
10	uf	UF do estabelecimento
11	municipio	Município
12	nu_telefone	Número de Telefone

Fig. 4. Colunas (em vermelho) excluídas do dataset do CNES.

No contexto deste trabalho, apenas são relevantes os registros de vacinação da primeira dose. Todos os registros de vacinação da segunda dose foram desconsiderados. O campo usado nesse filtro é o “Descrição Dose Vacina”.

Após a eliminação dos registros de vacinação que não serão analisados neste estudo, o dataset de vacinação ficou com 3.529.642 registros.

4.2 Tratamento dos dados / data cleaning

Identificação de valores nulos e/ou faltantes e definição de estratégia de tratamento (atribuição de valor ou eliminação da linha). Utilizando o Pandas, será usada uma máscara ([A-Z],[A-Z]) para identificar se os campos de UF, seja do estabelecimento, seja do vacinado estão no formato correto e, caso não estejam, a linha será eliminada, pois como o trabalho se propõe a analisar dados de uma região específica, havendo dúvidas se o dado pertence ou não ao conjunto, é melhor que se desconsidere para evitar ruídos no resultado. O mesmo será feito para sexo ([M|F]) e raça ([1|2|3|4|99]). No caso de raça, caso o valor presente seja diferente de “1-branca”, “2-preta”, “3-parda”, “4-amarela”, “99-sem informação”, será definido o valor para o campo como sendo “99-sem informação” (valor sentinela);

Identificação de registros duplicados e eliminação de um deles (caso os registros sejam exatamente iguais em todos os campos).

(revisar e completar seção com mais informações do tratamento de dados)

4.3 Proposta de modelo de extração de conhecimento e visualização dos dados

Nesta etapa, será utilizada a biblioteca matplotlib para que seja possível construir os gráficos definidos no item 3.2 anterior, utilizando os seus principais comandos (plot, show, grid e etc.).
(não podemos remover essa seção?)

4.4 Coleta de metadados da proveniência dos experimentos

Nesta etapa, serão coletadas as informações dos datasets que indicarão a origem dos dados, ou seja, de onde foram baixados, de que forma foram tratados e as transformações a que foram submetidos para análises futuras e posterior reuso.

Será usada a biblioteca PROV, uma implementação em Python do Modelo de Dados de Proveniência da W3C[9]. Ela suporta importação/exportação dos formatos PROV-O (RDF), PROV-XML e PROV-JSON.

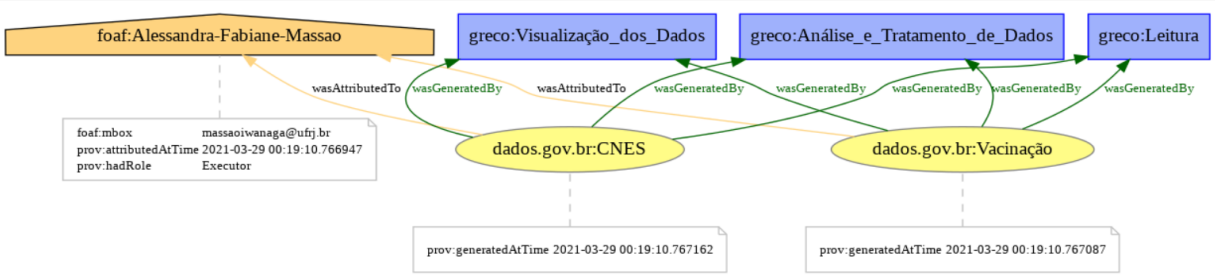


Fig. 5. Exemplo de grafo de proveniência gerado.

Como exemplificado na Figura 5, serão coletadas as seguintes informações de proveniência:

- O agente executor da atividade;
- O nome da atividade executada;
- O dataset consultado/consumido/analisado;
- Informações adicionais do agente executor:
  - E-mail;
  - Data e hora (timestamp) da atribuição do perfil ao agente;
  - Perfil do agente;
- Data e hora (timestamp) da geração dos dados de proveniência.

(revisar e atualizar a figura)

[9] World Wide Web Consortium (W3C) é uma comunidade internacional que desenvolve padrões abertos com o intuito de garantir o crescimento da Web a longo prazo. Link: https://www.w3.org/.

4.5 Projeto de reprodutibilidade do experimento

Um experimento reprodutível é aquele em que se é possível recriar o mesmo resultado usando código e dados originais, mesmo quando sendo feito por equipes, sistemas de medição e localidades diferentes [4]. Há de se atentar para não se confundir com o conceito de replicabilidade, que é a capacidade de recriar o mesmo resultado, usando, porém, novos dados, a partir do mesmo plano de experimento. Pode ocorrer do programa não ser reprodutível devido a diferentes razões, como: versão antiga do compilador, métodos ou funções descontinuados pela linguagem, erro de programação, etc.

Com o objetivo de se reproduzir o mesmo resultado deste trabalho, é necessário que se use o mesmo código, ambiente computacional e *datasets* usados no experimento. Como existe a possibilidade dos *datasets* serem modificados/atualizados em seus repositórios originários de armazenamento, foram disponibilizados publicamente para *download*, as versões dos *datasets* usadas na experimentação, assim como o código fonte do programa em linguagem Python.

Especificamente em relação a experimentos computacionais, garantir a reprodutibilidade por um longo período é um desafio, dada a evolução da tecnologia e a eventual descontinuidade de recursos utilizados. Nesse caso, busca-se descrever o ambiente computacional com o máximo de detalhe para que seja facilmente identificável qualquer ajuste necessário quanto não for possível utilizar exatamente o mesmo ambiente computacional.

Dessa forma, as seguintes medidas serão tomadas em relação ao código do experimento visando a reprodutibilidade do mesmo:

1. Identificação em detalhes do ambiente computacional.
2. Inclusão de comentários descrevendo as ações realizadas.
3. Utilização de funções ou recursos da versão mais atual do Python.
4. Tornar o código auto verificável.

5 Discussão

(essa é o capítulo mais necessário a ser aprimorado, nele vão encontrar tudo o que nos deparamos durante a implementação e apresentação dos resultados das análises obtidos. Talvez poderíamos criar um subcapítulo 4.6 dentro de Estudo de Caso apenas para exibir os resultados das análises e aqui discutir, mas eu acho que fica melhor botar os gráficos aqui, pois os prof. não iam gostar de colocarmos figuras no capítulo anterior sem discuti-las)

Havia sido planejado que para o campo raça/cor do paciente, caso o código estivesse diferente dos valores contidos no dicionário de dados, ou seja, diferente de “1”, “2”, “3”, “4” e “99”, que o valor seria alterado sumariamente para “99”, que significa “não especificado”. Porém, na implementação dessa limpeza, verificou-se que existiam registros com o código “5”, cujo nome da raça/cor era “indígena”. Diante desse dilema, decidimos não excluir essas linhas e considerar que houve uma falha na criação do dicionário de dados do dataset. Além do “5”, não houve nenhuma ocorrência de valor diferente dos listados.

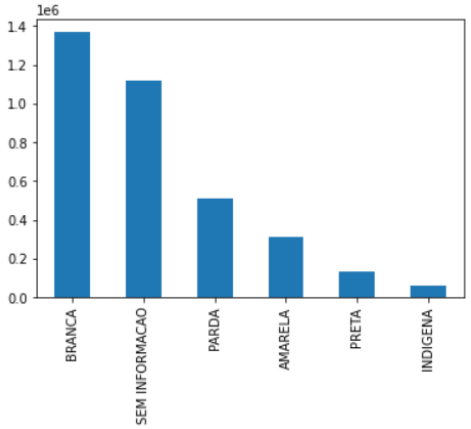


Fig. 5. Raça/cor dos vacinados.

Encontramos no dataset diversos casos de pacientes com mais de 2 registros de vacinação para a primeira dose da vacina, situação que não deveria acontecer. A decisão tomada com relação a esses casos foi de manter um único registro para cada paciente (paciente_id) com a menor data de vacinação, ou seja, o mais antigo foi mantido e o restante excluído.

	document_id	paciente_id	paciente_idade	paciente_dataNascimento	paciente_enumSexoBiologico	pa
4828100	56834bab-f0c1-486b-b309-db6076eb4b3b-i0b0	0c3390b4d0b9a535cb22799c2618053106a61deff30b85...	71.0	1949-06-14	M	99
3670249	4e09a541-fa26-4daa-8983-2ab84668a1e2-i0b0	0c3390b4d0b9a535cb22799c2618053106a61deff30b85...	71.0	1949-06-14	M	99
356327	ae5f43ff-7839-4851-b3f2-2183cb62d729-i0b0	0c3390b4d0b9a535cb22799c2618053106a61deff30b85...	71.0	1949-06-14	M	99
4773242	7f53f141-1d9c-4bcd-8d24-504b67d4ea50-i0b0	0fc334df6e1b34ea25366709ab808fd731e4f9f1c02347...	33.0	1987-08-15	M	99
1842150	11e92ee3-af7c-46c1-9096-e25235864e32-i0b0	0fc334df6e1b34ea25366709ab808fd731e4f9f1c02347...	33.0	1987-08-15	M	99
6830377	7d66212c-9a74-4ebe-ba99-4be3e2a79a43-i0b0	0fc334df6e1b34ea25366709ab808fd731e4f9f1c02347...	33.0	1987-08-15	M	99

Fig. 5. Registros com mais de um paciente_id.

Sobre a ocorrência de valores nulos, para as colunas X, Y, Z, nas ocorrências de valores nulos, a medida tomada foi…

Analisando a idade dos pacientes que tomaram a vacina, verificamos que a idade média é de 57 anos, a idade máxima é de 221 anos e a mínima é de 0 anos de vida. Notamos também que 1.376 vacinados têm menos de 18 anos de idade, o que contraria a recomendação de idade para a vacina. Além disso, 1.986 vacinados possuem idade maior que 110 anos.

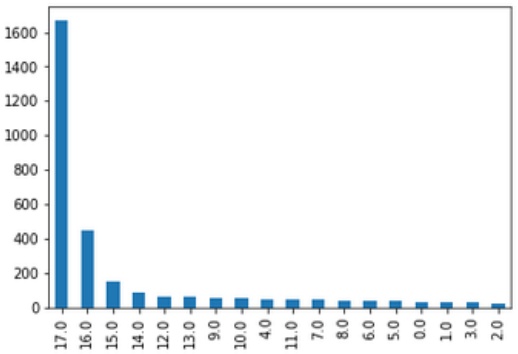


Fig. 5. Números de vacinados com menos de 18 anos de idade.

Para os casos de idade de vacinação abaixo de 18 anos ou acima de 110, a idade foi calculada a partir da data de nascimento, mas foram obtidos os mesmos valores dos originalmente informados no dataset, o que não permitiu qualquer ajuste em eventual erro de cadastro. Assim, esses valores de idade serão desconsiderados nas análises relacionadas à idade.

Verificou-se também que o sexo dos pacientes apresentou o valor “I”, além dos “F” de feminino e “M” de masculino. Como representavam menos de 2% do total de registros, ficou decidido mantê-los por serem irrelevantes no resultado final.

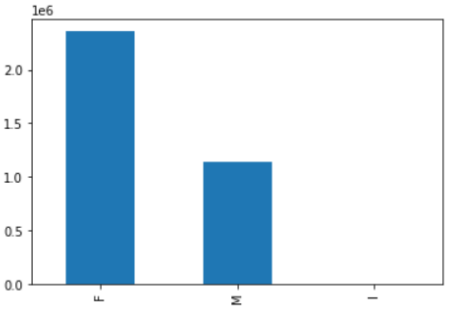
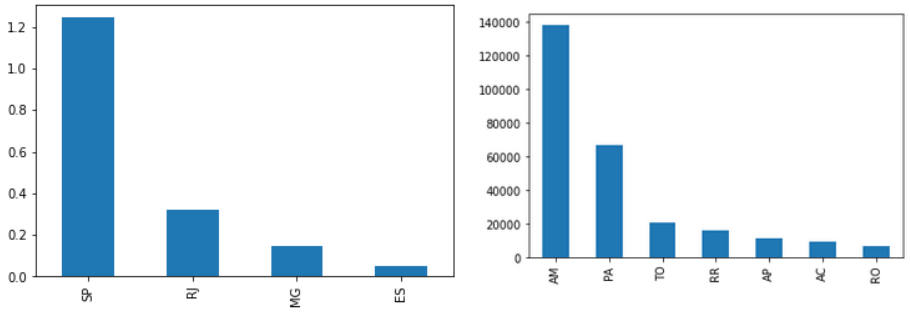


Fig. 5. Sexo biológico dos vacinados.

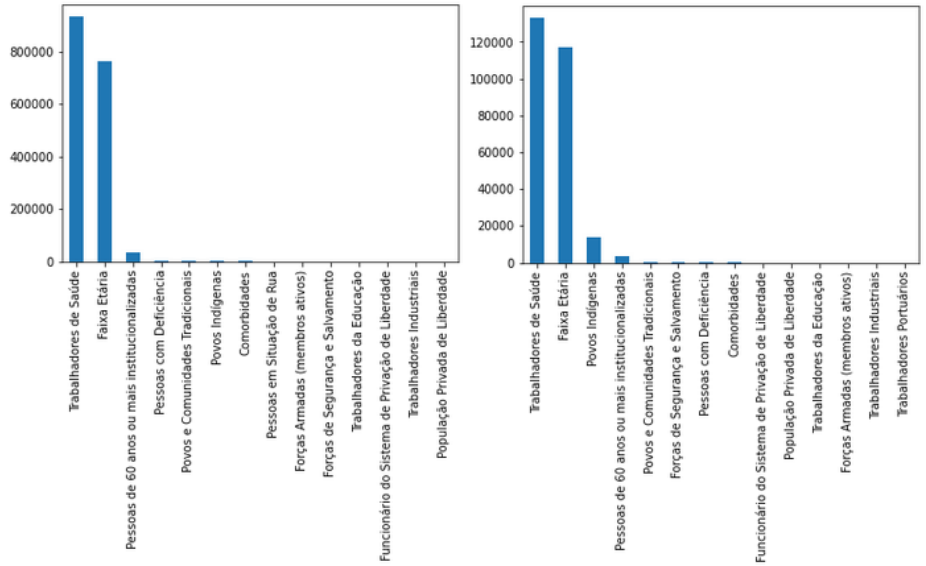
Foi observado que no dataset CNES ativos não foram encontrados registros duplicados e os nulos em um campo (qual?) irrelevante para a análise.

(abaixo estão os gráficos, à esquerda do sudeste e à direita do norte. a ideia é que o par seja apenas uma figura no artigo. Elaborar a redação que fale sobre esses resultados.)

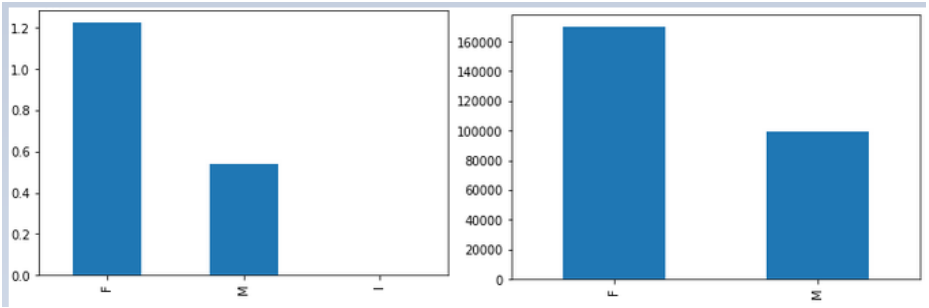
Quantitativo de vacinados nas regiões metropolitanas da região sudeste e norte por UF:



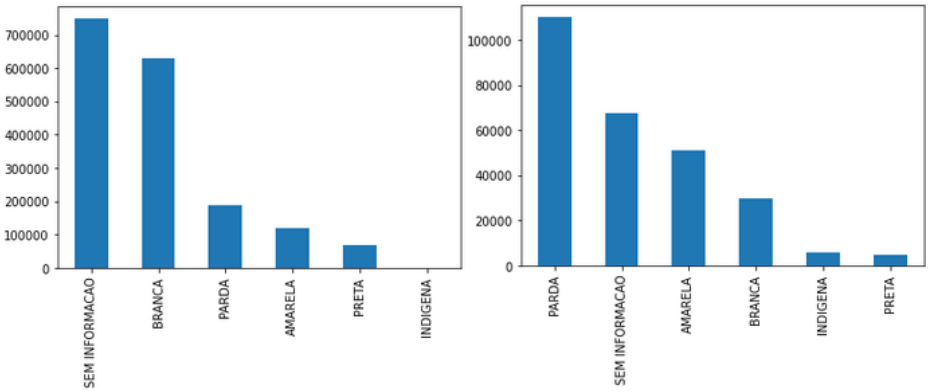
Quantitativo de vacinados nas regiões metropolitanas da região sudeste e norte por categoria:



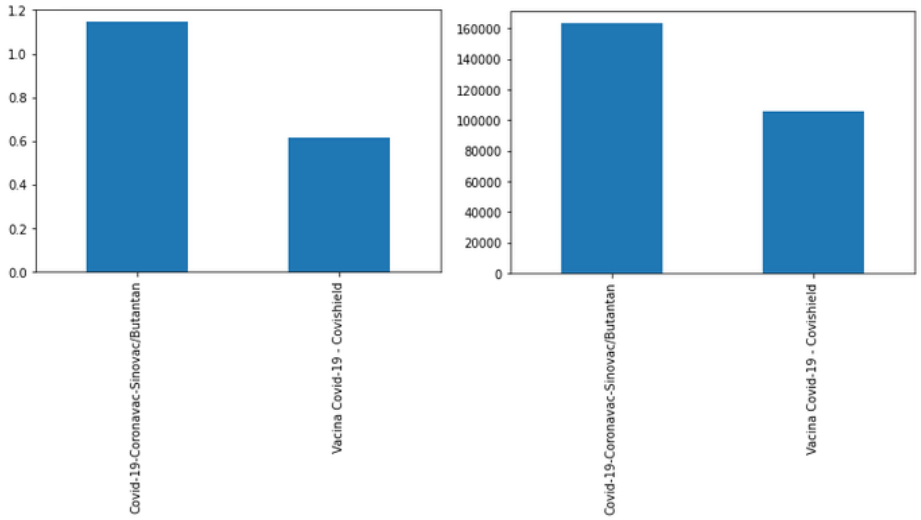
Quantitativo de vacinados nas regiões metropolitanas da região sudeste e norte por sexo biológico:



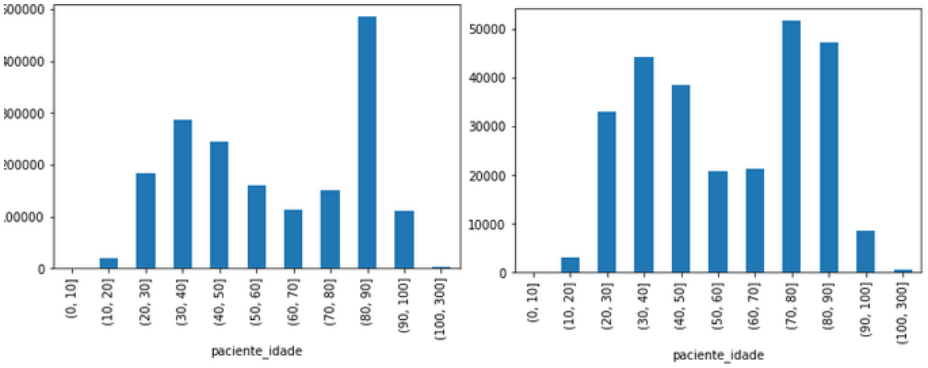
Quantitativo de vacinados nas regiões metropolitanas da região sudeste e norte por raça:



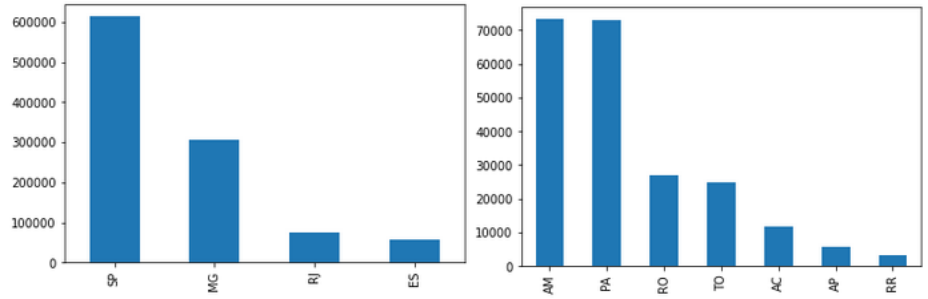
Quantitativo de vacinados nas regiões metropolitanas da região sudeste e norte por nome de vacina:



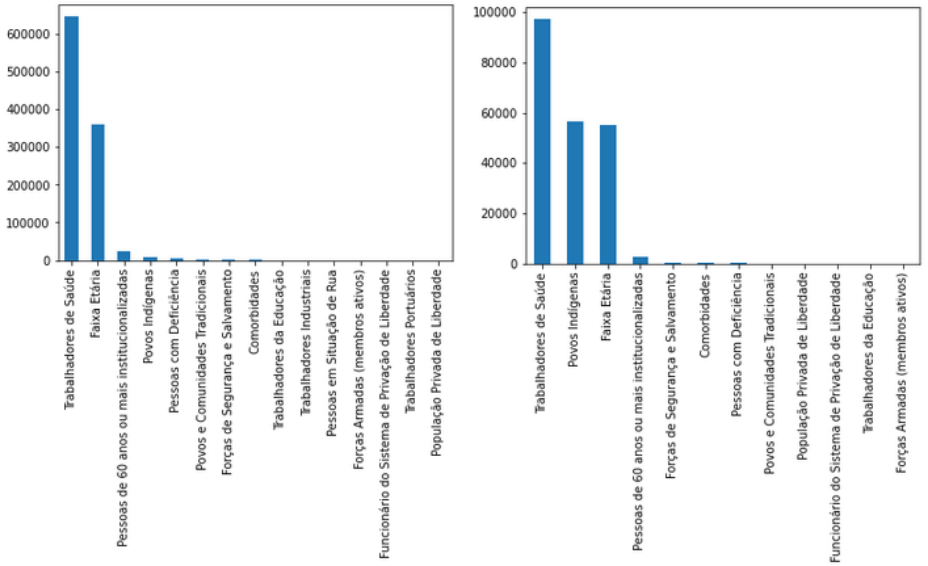
Quantitativo de vacinados nas regiões metropolitanas da região sudeste e norte por idade:



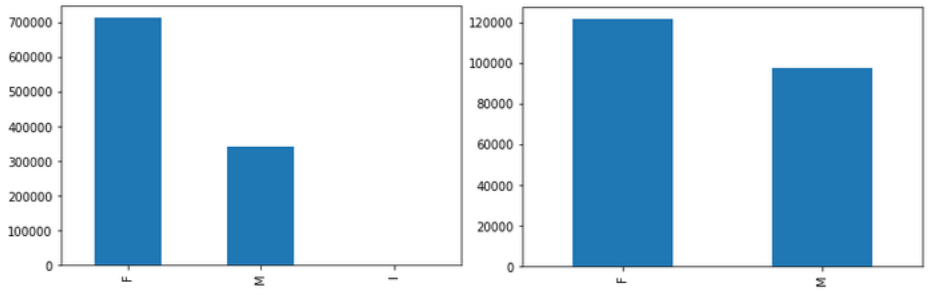
Quantitativo de vacinados nas regiões interioranas da região sudeste e norte por UF:



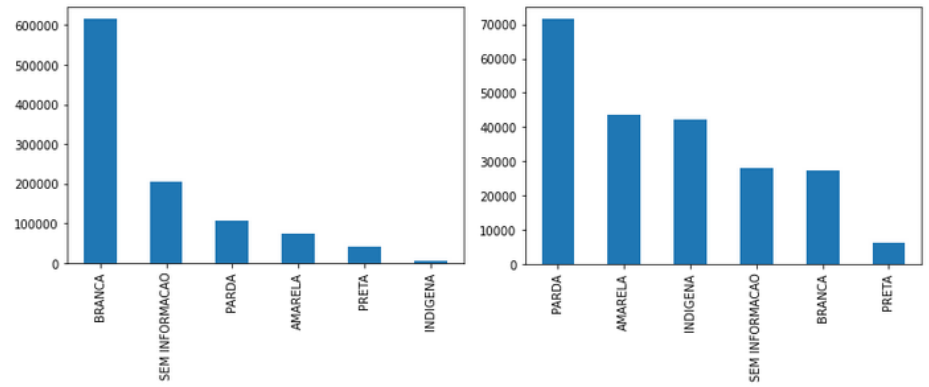
Quantitativo de vacinados nas regiões interioranas da região sudeste e norte por categoria:



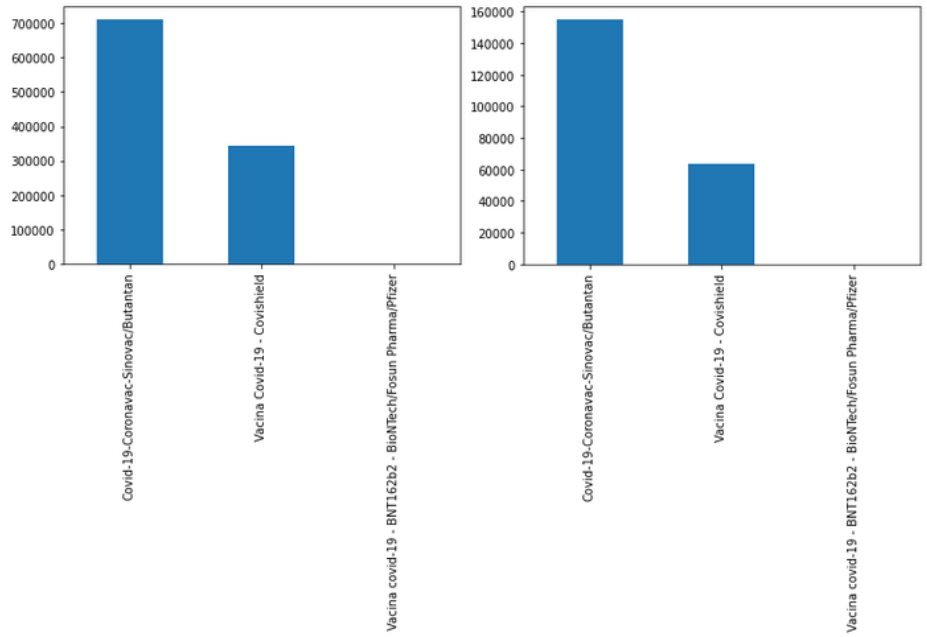
Quantitativo de vacinados nas regiões interioranas da região sudeste e norte por sexo biológico:



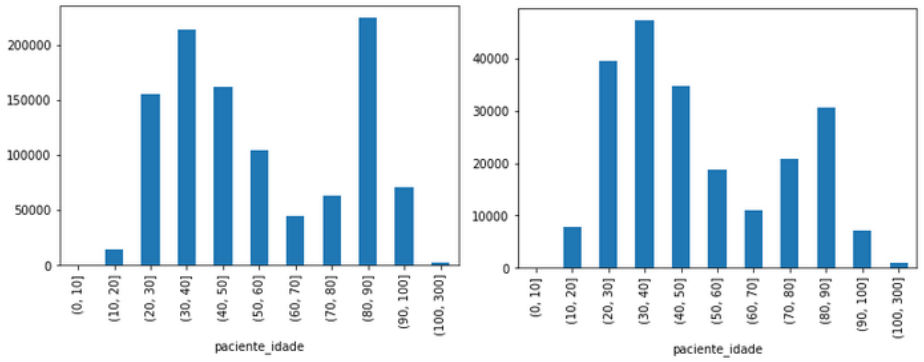
Quantitativo de vacinados nas regiões interioranas da região sudeste e norte por raça:



Quantitativo de vacinados nas regiões interioranas da região sudeste e norte por nome de vacina:



Quantitativo de vacinados nas regiões interioranas da região sudeste e norte por idade:



5.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either.

Subsequent paragraphs, however, are indented.

6 Conclusão

6.1 Resultado das Análises (os resultados das análises ficariam mais apropriadas no capítulo de discussão, na conclusão cabe mais um fechamento das ideias)

Com este trabalho, foi possível verificar as diferenças no perfil dos vacinados nas duas regiões, norte e sudeste, assim como nos estabelecimentos de saúde de vacinação dessas regiões.

Na região sudeste, o estado com mais vacinados foi o de São Paulo, seguido por Minas Gerais, Rio de Janeiro e Espírito Santo. Na categoria de vacinados, nas cinco primeiras posições ficaram os trabalhadores de saúde, seguidos por faixa etária, pessoas de 60 anos ou mais, povos indígenas e pessoas com deficiência. A maioria por sexo biológico é feminina. Por raça/cor, em primeiro lugar ficou a branca, seguidos de "sem informação", parda, amarela, preta e indígena. Sobre a vacina, a maioria foi vacinada com a Coronavac, seguida da Covishield (Oxford) e BioNtech/Pfizer. A maior parte da idade dos vacinados ficou na faixa dos 80 a 86 anos. Com relação aos estabelecimentos de saúde, a maioria deles é de gestão municipal, seguida de estadual e dupla (municipal e estadual). Os cinco tipos de estabelecimentos de saúde com mais vacinados foram: Centro de saúde/unidade básica, hospital geral, policlínica, posto de saúde e clínica/centro de especialidade.

Na região norte, o estado com mais vacinados foi o do Amazonas, seguido por Pará, Tocantins, Rondônia, Acre, Roraima e Amapá. Para a categoria de vacinados, nas cinco primeiras posições ficaram os trabalhadores de saúde, por faixa etária, povos indígenas, pessoas com 60 anos ou mais e força de segurança/salvamento. A maioria por sexo biológico é feminina. Por raça/cor, em primeiro lugar ficou a parda, seguido pelos "sem informação", amarela, branca, indígena e preta. Sobre a vacina, a maioria foi vacinada com a Coronavac, seguida pela Covishield e BioNtech/Pfizer. A maior parte das idades ficou na faixa de 80 a 83 anos. Sobre os estabelecimentos de saúde, a maioria é de gestão municipal, seguida pela estadual e dupla. Os cinco tipos de estabelecimentos de saúde que mais vacinaram foram: Centro de saúde/unidade básica, unidade de atenção à saúde indígena, hospital geral, unidade de vigilância em saúde e posto de saúde.

Abaixo encontra-se um breve resumo das análises encontradas em ordem decrescente:

Variáveis/Região:	Norte	Sudeste
Estados:	AM, PA,TO, RO, AC, RR, AP	SP, MG, RJ, ES
Categoria:	Trab. saúde, Faixa etária, Indígenas, 60 anos ou mais, Força Seg./Salv.	Trab. saúde, Faixa etária, 60 anos ou mais, Indígenas, Deficientes.
Sexo Biológico:	Feminino, Masculino	Feminino, Masculino
Raça/cor:	Parda, "Sem informação", Amarela, Branca, Indígena, preta	Branca, "Sem informação", Parda, Amarela, Preta, Indígena
Nome Vacina:	Coronavac, Covishield, BioNtech/Pfizer	Coronavac, Covishield, BioNtech/Pfizer
Idade:	80 a 83 anos	80 a 86 anos
Gestão Estab.:	Municipal, Estadual, Dupla	Municipal, Estadual, Dupla
Tipo Estab.:	Centro de saúde/Unidade básica, Unidade de atenção à saúde indígena, Hospital geral, Unidade de vigilância em saúde, Posto de saúde	Centro de saúde/Unidade básica, Hospital geral, Policlínica, Posto de saúde, Clínica/Centro de especialidade

Tabela 1. Resumo das análises por região..

……

6.2 Trabalhos Futuros (para um artigo, acho desnecessário esse subcapítulo, poderíamos colocar o texto abaixo direto na conclusão. Ou então alterar o título do capítulo para "Conclusão e trabalhos futuros")

Como trabalhos futuros, em primeiro lugar, como continuidade desta pesquisa, seria importante repetir as análises realizadas com dados mais avançados da vacinação no Brasil, com mais meses ou anos do seu início, para que se tenha um espectro maior de massa de dados. Com certeza, para este fim, será necessária uma infraestrutura mais robusta de ambiente de desenvolvimento que permita a importação de *datasets* com tamanhos grandes (maiores que 10 GB). Pode-se também expandir o escopo das análises para os imunizados, ou seja, os que já tomaram a segunda dose da vacina, inclusive. Nesse caso caberá a decisão de como lidar com casos de pacientes que tomaram a segunda dose de duas vacinas diferentes. Esse é um tema também interessante de ser aprofundado, pois foram encontrados muitos casos assim. Quanto ao recorte, faz-se necessário também analisar as demais regiões do Brasil, como nordeste, sul e centro-oeste.

incluir e elaborar esses pontos:

- Aprofundamento de análise sobre os registros duplicados e para identificar as causas dessa situação
- Avaliações sobre informações relativas às duas doses da vacina: mesma vacina e intervalo de tempo entre as doses

Referências

1. Ministério da Saúde. Cadastro Nacional de Estabelecimentos de Saúde (CNES) 2018. Disponível em: <https://dados.gov.br/dataset/cnes_ativo> Acesso em: 05 de março de 2021.
2. OpenDataSus. Campanha Nacional de Vacinação contra Covid-19 2021. Disponível em: <https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao> Acesso em: 05 de março de 2021.
3. Regiões Metropolitanas, Aglomerações Urbanas e Regiões Integradas de Desenvolvimento. Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/18354-regioes-metropolitanas-aglomeracoes-urbanas-e-regioes-integradas-de-desenvolvimento.html?=&t=acesso-ao-produto> Acesso em:07 de maio de 2021.

4. Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions Fabien C. Y. Benureau and Nicolas P. Rougier.
5. RIBEIRO, Marcelo Gomes. Território e Desigualdades de Renda em Regiões Metropolitanas do Brasil. **Dados**, Rio de Janeiro , v. 58, n. 4, p. 913-950, Dez. 2015 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0011-52582015000400913&lng=en&nrm=iso>. acesso em 07 Maio de 2021. http://dx.doi.org/10.1590/00115258201562.
6. GARNELO, Luiza; SAMPAIO, Sully. Bases sócio-culturais do controle social em saúde indígena: problemas e questões na Região Norte do Brasil. Cad. Saúde Pública, Rio de Janeiro , v. 19, n. 1, p. 311-317, Fev. 2003 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2003000100035&lng=en&nrm=iso>. Acesso em 07 Maio 2021. http://dx.doi.org/10.1590/S0102-311X2003000100035.
7. Donalisio, Maria Rita, Ruiz, Tânia, Cordeiro, Ricardo Fatores associados à vacinação contra influenza em idosos em município do Sudeste do Brasil. Revista de Saúde Pública [en linea]. 2006, 40(1), 115-119 [data de Consulta 7 de Mayo de 2021]. ISSN: 0034-8910. Disponível em: https://www.redalyc.org/articulo.oa?id=67240151018
8. LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.