UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

IGOR GONÇALVES GABARRA LOPES
Orientador: Prof. Dr. Rafael Alves Bonfim de Queiroz
Coorientador: Prof. Dr. Anderson Almeida Ferreira
Coorientadora: Profa. Dra. Cláudia Martins Carneiro

ANÁLISE DO IMPACTO DA PANDEMIA DA COVID-19 NA REALIZAÇÃO DE EXAMES DE PAPANICOLAU NO BRASIL

Ouro Preto, MG
2023

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

IGOR GONÇALVES GABARRA LOPES

ANÁLISE DO IMPACTO DA PANDEMIA DA COVID-19 NA REALIZAÇÃO DE EXAMES DE PAPANICOLAU NO BRASIL

Monografia I I apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Rafael Alves Bonfim de Queiroz
Coorientador: Prof. Dr. Anderson Almeida Ferreira
Coorientadora: Profa. Dra. Cláudia Martins Carneiro

Ouro Preto, MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Igor Gonçalves Gabarra Lopes

Análise do impacto da pandemia do Covid-19 na realização de exames de Papanicolau no Brasil

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 20 de Março de 2023.

Membros da banca

Rafael Alves Bonfim de Queiroz (Orientador) - Doutor - Universidade Federal de Ouro Preto
Anderson Almeida Ferreira (Coorientador) - Doutor - Universidade Federal de Ouro Preto
Cláudia Martins Carneiro (Coorientadora) - Doutora - Universidade Federal de Ouro Preto
Fernanda Sumika Hojo de Souza (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Mariana Trevisan Rezende (Examinadora) - Doutora - Universidade Federal de Ouro Preto

Rafael Alves Bonfim de Queiroz, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 20/03/2023.

Documento assinado eletronicamente por **Rafael Alves Bonfim de Queiroz**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 29/03/2023, às 20:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0493035** e o código CRC **61DC9A65**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho à minha mãe e irmã que me apoiam e motivam todos os dias, sem exceção.

Agradecimentos

Agradeço primeiramente à minha mãe Tereza, minha irmã Luiza e minha avó Elcy por todo o apoio, suporte e incentivo durante toda a vida. Em especial, agradeço ao meu avô Fernando, que mesmo não estando mais entre nós, me inspira como exemplo de ser humano.

Agradeço ao meu orientador, Rafael Alves Bonfim de Queiroz, pelo incentivo, apoio e orientação durante toda a elaboração do presente trabalho. Além disso, agradeço ao coorientador Anderson Almeida Ferreira por aceitar fazer parte do projeto a partir da Monografia II, agregando ao trabalho todo o seu conhecimento obtido ao longo de sua carreira. Agradeço também a coorientadora Cláudia Martins Carneiro por ter apresentado o tema desta monografia ao meu orientador.

Aos professores do curso de Ciência da Computação da UFOP, agradeço pelo profissionalismo e excelência ao transmitir todo o conteúdo necessário para a minha formação. Em especial, gostaria de agradecer o professor Rodrigo Ribeiro por me guiar durante a Iniciação Científica e ao professor Joubert Lima, que além de toda inspiração pessoal e profissional, aceitou fazer parte de um projeto que leva a computação para jovens alunos de uma escola pública.

Aos meus colegas de curso: Ana, Gustavo, Marcelo, Paula, Sammuel, Valmir e Vivyann, agradeço pelo convívio, companheirismo e troca durante toda a graduação. Em especial, agradeço ao Alexsandro pela amizade, compreensão e inspiração por ser um profissional, amigo e pai exemplar.

Aos amigos de longa data, Augusto, Bruno, Fábio Daniel, Leonardo, Lucas Oliveira e Rodolpho, muito obrigado pelo convívio, apoio, aprendizado e amizade durante todos esses anos de vida.

Agradeço à empresa Efí pelo ambiente ideal para crescimento e evolução profissional. Minha gratidão aos colaboradores Dayvis Apolinário, João Souza, Lucas Souza, Marcelo Nepomuceno e Marcos Alvarenga pelos ensinamentos diários, companheirismo e inspiração profissional. Em especial, gostaria de agradecer ao colega de trabalho Lucas Cassimiro, que felizmente acabou se tornando um grande amigo para a vida.

Por fim, muito obrigado a todos que já me incentivaram ou forneceram suporte durante todos esses anos, com certeza tais atitudes foram fundamentais para me tornar o que sou hoje.

Resumo

O câncer do colo do útero é causado quando ocorre uma infecção persistente de certos tipos do papilomavírus humano (HPV), ocorrendo o desenvolvimento de alterações celulares que podem ser identificadas a partir do exame de Papanicolau. A realização periódica destes exames é importante método preventivo. Porém, devido ao surgimento da Covid-19, uma infecção respiratória, potencialmente grave e com altas taxas de transmissão, recomendou-se medidas restritivas para a população, incluindo evitar a realização de exames de rotina. O presente estudo tem como objetivo demonstrar o impacto da pandemia da Covid-19 na realização de exames de Papanicolau e seus diferentes tipos de resultados em todo o território nacional. Para isso, foi desenvolvido um sistema utilizando a linguagem Javascript em conjunto com o Node.Js, com o objetivo de processar as informações presentes no SISCAN do DATASUS, uniformizando os diferentes tipos de informações presentes na base de dados, persistindo-as em um banco de dados relacional PostgreSQL e, por fim, elaborando um conjunto de visualizações gráficas com dados temporais entre os anos de 2013 e 2022, utilizando a linguagem Python em conjunto com as bibliotecas Pandas e Seaborn. Os resultados indicaram que a maioria das regiões e estados brasileiros apresentaram queda significativa do número de exames no primeiro ano de pandemia (2020), com aumento gradativo nos dois anos seguintes onde, em alguns casos, os dados de 2022 chegaram a ultrapassar as informações referentes ao último ano anterior ao início do período pandêmico, em 2019.

Palavras-chave: SISCAN. Câncer de colo do útero. Covid-19. Pandemia.

Abstract

Cervical cancer is caused when there is a persistent infection of certain types of human papillomavirus (HPV), resulting in the development of cellular changes that can be identified from the Pap test. The periodic follow-up of these exams is an important preventive method. However, due to the emergence of the Covid-19, a potentially severe respiratory infection with high transmission rates, restrictive measures were recommended for the population, including avoiding routine exams. The present study aims to demonstrate the impact of the Covid-19 pandemic on the performance of Papanicolaou exams and their different types of results throughout the national territory. For this, a system was developed using Javascript language with Node.Js framework to process the information present in the SISCAN of the DATASUS, standardizing different types of information present in the dataset, persisting them in a relational database named PostgreSQL and, finally, elaborating sets of graphical visualizations with temporal data between the years of 2013 and 2022, using Python language with Pandas and Seaborn libraries. The results indicated that most Brazilian regions and states presented a significant drop in the number of exams in the first year of the pandemic (2020), with a gradual increase in the following two years where, in some cases, the data for 2022 even exceeded the information referring to to the last year before the beginning of the pandemic period, in 2019.

Keywords: SISCAN, Cervical Cancer, Covid-19, Pandemic.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

API	*Application Programming Interface*, ou Interface de Programação de Aplicações
Covid-19	Coronavirus Disease - 2019
CSV	*Comma-separated Values*, ou Valores Separados por Vírgulas
DATASUS	Departamento de Informática do Sistema Único de Saúde do Brasil
ETL	*Extract, Transform and Load* (Extrair, Transformar e Carregar)
HPV	*Human Papillomavirus* (Papilomavírus humano)
RDBMS	*Relational Database Management System* (Sistema de Gerenciamento de Banco de Dados Relacional)
SGBD	Sistema de Gerenciamento de Banco de Dados
SISCAN	Sistema de Informação do Câncer
SISCOLO	Sistema de Informação do Câncer de Colo do Útero
SQL	*Structured Query Language* (Linguagem de Consulta Estruturada)
SUS	Sistema Único de Saúde
UF	Unidade Federativa

Sumário

1 Introdução

De acordo com o Instituto Nacional de Câncer (INCA, 2022), o câncer do colo do útero é causado pela infecção persistente por alguns tipos do Papilomavírus Humano - o HPV. É uma infecção frequente e na maioria dos casos não causa doença. Porém, em alguns casos, ocorre a alteração celular, o que acaba evoluindo para o câncer. Tais alterações podem ser identificadas em um exame preventivo e periódico, o exame Papanicolau. Em suma maioria, o câncer em questão é curável quando a identificação de alterações genéticas ocorre em seu estágio inicial. Sendo assim, é identificada a importância da realização periódica do exame Papanicolau como método preventivo. Porém, devido ao surgimento da pandemia do coronavírus e a diminuição do convívio social durante os principais anos pandêmicos, pode ter impactado na realização deste exame. Desta forma, é revelante analisar as informações para verificar a situação dos exames realizados em todo o Brasil.

O coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2), ou Covid-19, é o sétimo coronavírus humano e foi descoberto em Wuhan, na China, durante uma epidemia de pneumonia em janeiro de 2020. A partir disso, o vírus se difundiu no mundo e até o fim de maio do mesmo ano, foram identificados aproximadamente 308 mil mortos e 4,8 milhões de pacientes infectados pela doença (CIOTTI et al., 2020).

Por ser um vírus do grupo dos coronavírus, a Covid-19 é uma doença altamente contagiosa. Seu principal método de contágio é o contato respiratório entre um paciente infectado e um não infectado (WHO, 2021), o que justifica a criação de políticas públicas para promover o distanciamento social, evitar a infecção desenfreada da população e tentar esquivar-se de um colapso no sistema de saúde de maneira mundial. Dentre as principais políticas públicas adotadas mundialmente antes da criação de vacinas, podemos citar como mais eficazes a execução de *lockdowns* eficientes, utilização de máscaras em ambientes, sejam eles abertos e fechados e a limpeza frequente das mãos.

Para uma execução eficiente de um *lockdown*, é esperado que a população tenha consciência de que o convívio social deve ser evitado, exceto em emergências. Por isso, foram observadas quedas abruptas em alguns setores da economia: entretenimento, lazer, moda, entre outros. Além disso, notou-se essa queda em alguns âmbitos da saúde como, por exemplo, a diminuição na disponibilidade de médicos para realização de consultas, acompanhamentos e cirurgias em áreas que necessitam de menor urgência e a queda na realização de exames de rotina para prevenção de alguns tipos de doença que evoluem ao longo do tempo, tendo como principal o câncer. Dentre alguns desses tipos de câncer, temos o câncer do colo do útero, no qual será avaliado no escopo deste trabalho.

O presente trabalho teve como objetivo avaliar as informações disponíveis no SISCAN

(Sistema de Informação do Câncer) do DATASUS (Departamento de Informática do Sistema Único de Saúde do Brasil), este com dados entre 2013 e 2022, no que relaciona-se a realização de exames de Papanicolau e identificação de tipos de resultados do exame, denominados como indicadores, avaliando quantitativamente as informações anteriores e posteriores ao surgimento da pandemia da Covid-19. Para tal, foi desenvolvido um sistema que extrai, processa e persiste os dados do SISCAN em um banco de dados relacional para que, com isso, seja possível produzir um sistema de visualização de dados com diversos tipos de atributos e períodos distintos, com a finalidade de obter *insights* sobre aquilo que foi produzido. Para extração tratamento dos dados, utiliza-se a linguagem Javascript e seu *framework* Node.JS e, para persistência dos dados, adota-se o banco de dados relacional PostgreSQL com a execução de consultas em SQL para geração de valor a partir dos dados coletados. Para processamento dos resultados das consultas, emprega-se a linguagem *Python* em conjunto com as bibliotecas *Pandas* para modelagem de dados e *Seaborn* para visualização de dados.

1.1 Problema abordado

O problema a ser abordado no presente trabalho é quantificar o impacto da pandemia da Covid-19 no que tange a informações relacionadas ao câncer do colo do útero, identificando:

- A situação das informações relacionadas ao câncer do colo do útero anteriores a pandemia;
- A tendência que os dados indicam para os anos pandêmicos;
- A diferença entre os dados consolidados e a tendência obtida durante o período.

1.2 Justificativa

A elaboração do presente trabalho se justifica por três principais motivos: (1) necessidade de visualização temporal das informações relacionadas aos exames de Papanicolau durante a pandemia da Covid-19; (2) necessidade de visualização de mais de um atributo disponível no sistema construído para visualização dos dados do SISCAN/DATASUS; (3) poucos trabalhos foram elaborados com objetivo de efetuar uma análise de dados sobre a queda de exames de citologia do colo do útero durante o período da pandemia, analisando todas as informações disponíveis no sistema, entre 2013 e 2022.

1.3 Objetivos geral e específicos

Este trabalho tem como objetivo geral analisar o impacto da pandemia da Covid-19 na realização de exames de Papanicolau no Brasil. Para isso, adotaram-se dados do SISCAN/DA-

TASUS separados por ano, unidade federativa (UF) de residência e indicador. São objetivos específicos:

- Analisar dados do SISCAN sobre informações temporais relacionadas ao câncer de colo do útero anteriores (2013 a 2019) e durante a pandemia da Covid-19 (2020 a 2022);
- Aplicar ferramentas para extração, tratamento, persistência e visualização dos dados obtidos;
- Quantificar o impacto da pandemia da Covid-19 na realização de exames de Papanicolau, elaborando gráficos que representam informações entre os anos de 2013 e 2022.

1.4 Organização da monografia

Os demais capítulos estão organizados como seguem:

Capítulo 2: expõem-se os conceitos utilizados para a elaboração da presente monografia e apresentam-se alguns trabalhos da literatura relacionados ao tema aqui tratado;

Capítulo 3: detalha-se o processo de construção do sistema proposto neste documento;

Capítulo 4: apresentam-se os resultados obtidos para o presente trabalho;

Capítulo 5: sintetizam-se as conclusões e trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo, é apresentado a fundamentação teórica, com conceitos importantes para o entendimento do trabalho realizado, dividido em cinco seções: Javascript e Node.Js (Seção 2.1), SQL e PostgreSQL (Seção 2.2), bibliotecas Pandas e Seaborn do Python (Seção 2.3), exame de Papanicolau (Seção 2.4) e SISCAN (Seção 2.5) . As duas primeiras seções apresentam as definições e casos de usos. A terceira seção descreve a linguagem de programação Python e o ecossistema para plotagem de gráficos utilizando as bibliotecas Pandas e Seaborn. Na quarta seção explica brevemente o exame de Papanicolau. Por fim, a Seção 2.5 explica em linhas gerais o SISCAN e detalham-se os resultados possíveis para um exame e lista os indicadores inclusos em cada resultado.

2.1 Javascript e Node.JS

JavaScript é uma linguagem de programação de interpretação de script consolidada em toda a Web. Foi criado em 1995 (WIRFS-BROCK; EICH, 2020) com o principal objetivo de promover interação em páginas da internet. Atualmente, pode ser utilizada em praticamente qualquer tipo de aplicação, incluindo aquelas que propõem a manipulação de dados para geração de valor, o que é o caso do presente documento.

O Node.JS é uma plataforma de código aberto que permite a execução de código JavaScript em aplicações no lado do servidor. Tem como objetivo a construção de aplicações robustas e altamente escaláveis. Com ele, é possível criar conexões com diferentes bancos de dados e manipular suas informações da maneira que desejar.

Como um dos objetivos deste trabalho é o desenvolvimento de um sistema que seja capaz de processar informações heterogêneas, realizar transformações e comunicar-se com um banco de dados, a linguagem Javascript combinada com o Node.JS foi extremamente necessária por facilitar o processo de desenvolvimento do sistema proposto neste trabalho. O emprego destas ferramentas acabou fornecendo agilidade por possibilitar a utilização de pacotes criados e mantidos constantemente pela comunidade, além de permitir a conexão direta com um banco de dados relacional de maneira simples e direta.

2.2 SQL e PostgreSQL

O SQL (ou linguagem de consultas estruturada) é uma linguagem de consulta universal utilizada para o gerenciamento de bancos de dados relacionais. Foi criado em 1970 e é utilizado para gerenciar e armazenar grandes quantidades de dados (OPPEL; SHELDON, 2008). Para

utilizar o SQL, é necessária a presença de um Sistema de Gerenciamento de Banco de Dados (SGBD). Com ele, é possível inserir, deletar, ler e atualizar registros salvos em sua memória, com a possibilidade de extrair informações e gerar valor a partir de dados. Dentre os principais bancos de dados relacionais disponíveis atualmente, temos o PostgreSQL.

O PostgreSQL é um SGBD de código aberto e utilizado em bancos de dados relacionais. Foi criado em 1996 e tem como objetivo armazenar, gerenciar e recuperar informações disponíveis em formato de tabelas heterogêneas.

A utilização da linguagem SQL em conjunto com o PostgreSQL trouxe grande contribuição para elaboração do presente trabalho, principalmente pela facilidade de instalação, configuração e utilização do PostgreSQL em conjunto com a possibilidade de gerar consultas que envolvem determinado tipo de inteligência para filtragem e cruzamento das informações, evitando, assim, a verificação manual dos dados.

2.3 Bibliotecas Pandas e Seaborn do Python

O Python é uma linguagem de programação interpretada, imperativa, com tipagem dinâmica e suporte aos paradigmas funcional e orientada a objetos. Possui diversas bibliotecas gratuitas e de código aberto (*open-source*) nas quais têm sido empregadas em diversos projetos nas áreas de ciência e análise de dados, inteligência artificial e aprendizado de máquina. Pode ser considerada como uma das principais linguagens a serem adotadas pela comunidade para manipulação de dados, e alguns fatores elevaram a sua popularidade:

- Linguagem de código aberto;
- Facilmente escalável;
- Uma comunidade ativa, o que nos proporciona uma grande quantidade de bibliotecas específicas e atualizadas. Em destaque, a biblioteca Pandas é usada para manipulação e análise de dados, possibilitando ler, manipular e plotar dados.

Estas funcionalidades possibilitaram o início da análise exploratória de dados, ou seja, a mineração dos dados. Além disso, a visualização de dados possibilita a exibição gráfica de um grupo de informações, inicialmente sem sentido, mas que podem possuir características que muitas vezes são visíveis somente quando distribuídas espacialmente.

2.3.1 Pandas

O Pandas (MCKINNEY et al., 2011) é um pacote disponível para a linguagem Python que possibilita o trabalho com dados tabelados de maneira simples e intuitiva, com o objetivo de ser a principal ferramenta para análise de dados práticos e do mundo real. Além disso, o pacote

almeja ser a maior, mais poderosa e mais flexível ferramenta de código aberto para qualquer linguagem de programação. Geralmente, o Pandas é indicado para diversos agrupamentos de informações. Seguem alguns exemplos:

- Dados em tabelas com colunas de tipos heterogêneos;
- Dados de séries temporais estejam elas ordenadas ou não;
- Matrizes homogêneas ou heterogêneas com rótulos definidos;
- Qualquer tipo de conjunto de dados estatístico ou observacional.

2.3.2 Seaborn

O Seaborn (WASKOM, 2021b) é uma biblioteca de visualização de dados da linguagem Python, baseada em uma outra biblioteca, a Matplotlib. Tem como principal objetivo a elaboração de gráficos estatísticos de maneira fácil, simples e atrativa ao usuário para explorar e entender o que significam os dados utilizados. Sua biblioteca possui funções declarativas que permitem trazer o foco no significado dos dados ao invés de deslocar esforços para identificar e construir maneiras para visualizá-los.

Na Figura 2.1, é demonstrado um exemplo de código-fonte com o objetivo de mostrar o funcionamento da plotagem de gráficos utilizando o Seaborn:

```
1  import seaborn as sns
2
3  # Aplicação do tema básico da biblioteca
4  sns.set_theme()
5
6  # Carregamento de um dataset
7  tips = sns.load_dataset("tips")
8
9  # Criação de um gráfico utilizando o dataset carregado
10 sns.relplot(
11     data=tips,
12     x="total_bill", y="tip", col="time",
13     hue="smoker", style="smoker", size="size",
14 )
```

Figura 2.1 – Exemplo de codificação para realizar uma visualização utilizando o Seaborn (WASKOM, 2021a).

No exemplo da Figura 2.1, temos um trecho de código responsável por criar um gráfico referente ao conjunto de dados denominado de "tips". Para plotagem do gráfico em si, foi necessário somente a realização da chamada da função *relplot*, onde forneceu-se somente a

relação de cada variável no gráfico em relação as colunas presentes no conjunto de dados. A partir disso, o gráfico da Figura 2.2 foi elaborado.

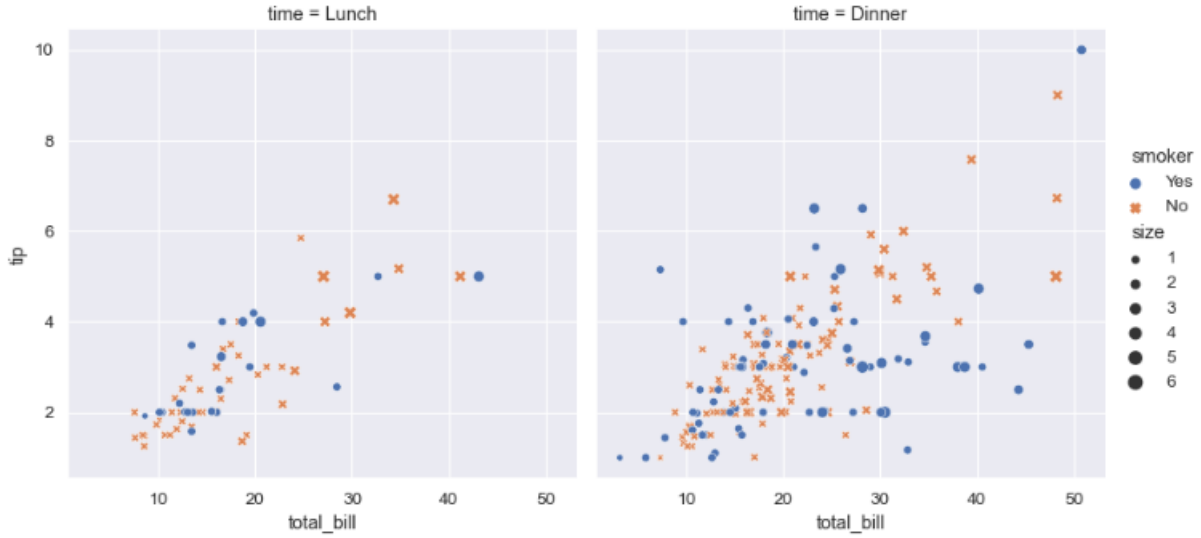


Figura 2.2 – Gráfico montado utilizando a função relplot do Seaborn (WASKOM, 2021a).

No Seaborn (WASKOM, 2021b) todas as suas funções são acessíveis de maneira pública e, apesar de toda a sua implementação ser estruturada hierarquicamente, duas funções de módulos distintos podem produzir praticamente a mesma visualização. Os módulos são denominados de: relacionais, categóricos e distribucionais. As funções dentro de um módulo podem compartilhar recursos que não estão disponíveis em outros, com o objetivo de facilitar a alternância entre tipos de representações enquanto realiza o processo de exploração de dados, visto que cada tipo de representação pode diferenciar a história na qual o analista quer contar ao usuário.

As funções podem ser classificadas como funções a nível de figura ou a nível de eixo. As funções a nível de eixo plotam os dados em um objeto, enquanto as funções a nível de figura gerenciam o objeto. Cada módulo possui uma única função a nível de figura e existem um conjunto de funções a nível de eixo disponíveis para sua utilização. A organização segue a estrutura ilustrada na Figura 2.3.

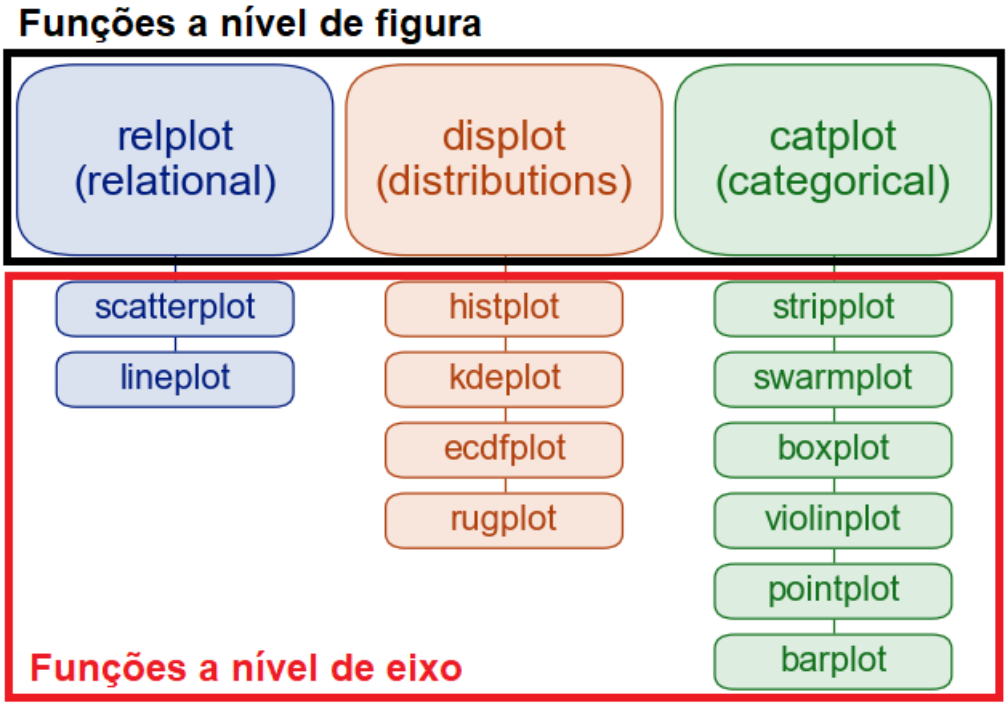


Figura 2.3 – Organização das funções a nível de eixo e de figura. Fonte: documentação do *seaborn* (WASKOM, 2021a).

Por exemplo, a função *relplot* é a função a nível de figura para o módulo relacional. Seu comportamento padrão é a elaboração do gráfico do primeiro elemento hierárquico nas funções a nível de eixo, o gráfico de dispersão (*scatterplot*). Para desenhar outra função a nível de eixo, deve-se manter a mesma chamada da função com o acréscimo do parâmetro *kind*. Nesse caso, para desenhar um gráfico de linha, deve-se incluir *kind="line"*.

2.4 Exame de Papanicolau

De acordo com o Ministério da Saúde do Brasil (BRASIL, 2011), o exame de Papanicolau (ou exame preventivo de colo do útero) é um procedimento com o objetivo de detectar alterações celulares no colo do útero. Sua principal estratégia é a detecção dessas alterações de maneira precoce para diagnóstico da doença ainda em fase inicial, antes mesmo que a mulher apresente qualquer tipo de sintoma. De acordo com (SILVA, 2016), deve ser realizado periodicamente em todas as mulheres que tem ou já tiveram vida sexual ativa e que possuem entre 25 e 64 anos.

Todo resultado de exame de Papanicolau realizado deve ter um resultado vinculado a si mesmo. Para cada resultado, é necessária a presença de pelo menos um indicador, com o objetivo de demonstrar a análise realizada em relação ao exame. A seguir, serão demonstrados e explicados cada tipo de indicador considerado e os grupos nos quais cada indicador pode estar incluso.

2.5 SISCAN

O SISCAN é um sistema gerido pelo Ministério da Saúde através da Secretaria de Assistência à Saúde em conjunto com as Secretarias Estaduais de Saúde e as Secretarias Municipais de Saúde. Tem como principal objetivo a disponibilização pública de dados para tabulação, tendo como um dos exemplos de dados disponíveis as informações sobre o câncer do colo do útero. Os dados são processados e disponibilizados pelo DATASUS - Departamento de Informática do SUS - da Secretaria Executiva do Ministério da Saúde. No SISCAN são registrados dados relacionados aos exames de Papanicolau conforme apresentados a seguir.

Os dados abaixo descritos possuem determinado indicador, o qual determina uma característica encontrada no exame realizado (DATASUS, 2019). Tais indicadores formam grupos de resultados. De maneira geral, um tipo de resultado deste grupo indica qual é o diagnóstico para determinado exame. Temos, principalmente, o tipo de medida denominado de "Quantidade de exames", onde todos os exames realizados em um determinado período estão inclusos. Além disso, temos os resultados negativos (indicando a ausência de alterações celulares pré-malignas e malignas) e resultados alterados (indicando a ausência de alterações celulares pré-malignas e malignas). Dentre os resultados alterados, é possível distingui-los, também, entre resultados alterados glandulares (de acordo com (ONCOGUIA, 2014), quando células glandulares não se parecem normais mas apresentam características cancerígenas) ou resultados alterados escamosos (de acordo com (TELESSAúDERS, 2015), quando ocorre lesões na mucosa do colo com exposição do estroma). Abaixo, encontram-se a lista de indicadores que compõem cada tipo de resultado acerca de um exame realizado:

- Para aqueles resultados identificados como resultados negativos, temos:
  - *Alt. ben: Inflamação:* exame negativo para neoplasia do colo do útero com alterações celulares benignas reativas ou reparativas associadas à inflamação;
  - *Alt. ben: Metap. Esca:* exame negativo para neoplasia do colo do útero com alterações celulares benignas reativas ou reparativas associadas a metaplasia escamosa imatura;
  - *Alt. ben: Reparação:* exame negativo para neoplasia do colo do útero com alterações celulares benignas reativas ou reparativas associadas à reparação;
  - *Alt. ben: Atr. Inflam:* exame negativo para neoplasia do colo do útero com alterações celulares benignas reativas ou reparativas associadas à atrofia com inflamação;
  - *Alt. ben: Radiação:* exame negativo para neoplasia do colo do útero com alterações celulares benignas reativas ou reparativas associadas à radiação;
  - *Alt. ben: Outros:* exame negativo para neoplasia do colo do útero com alterações celulares benignas reativas ou reparativas associadas a outras causas.
- Para aqueles resultados identificados como resultados alterados, temos:

  - *Ori. Indef. Não Neo:* presença de células atípicas de origem indefinida, possivelmente não neoplásicos;
  - *Ori. Indef. Alto Grau:* presença de células atípicas de origem indefinida onde não se pode afastar lesão de alto grau.
  - *Outras Neoplasias:* diagnóstico referente a outras neoplasias malignas no colo do útero.
  - Resultados Alterados Escamosos:
    * *ASC-US:* presença de células escamosas atípicas de significado indeterminado possivelmente não neoplásicas;
    * *ASC-H:* presença de células escamosas atípicas de significado indeterminado onde não se pode afastar lesão de alto grau;
    * *Les IE Baixo Grau:* presença em células escamosas de lesão intra-epiteliais de baixo grau - compreendendo efeito citopático pelo HPV e neoplasia intra-epitelial cervical grau I;
    * *Les IEp Alto Grau:* presença em células escamosas de lesão intra-epitelial de Alto grau – compreendendo neoplasia intra-epiteliais cervicais graus II e III;
    * *Les IE AG Mic. Inv:* presença em células escamosas de lesão intra-epiteliais de Alto grau, não podendo excluir micro-invasão;
    * *Carc. Epiderm. Inv:* diagnóstico de Carcinoma epidermoide invasor.
  - Resultados Alterados Glandulares:
    * *At. Glan. Ind. Não Neo:* presença de células glandulares atípicas de significado indeterminado possivelmente não neoplásicas;
    * *At. Glan. Ind. Alto Grau:* presença de células glandulares atípicas de significado indeterminado onde não se pode afastar lesão de alto grau;
    * *Adenocarc in situ:* diagnóstico de Adenocarcinoma *in situ*;
    * *Adenocarc invasor:* diagnóstico descritivo de Adenocarcinoma invasor. Compreende laudos cervical, endometrial ou sem outras especificações.

- Além disso, existem alguns outros tipos de indicadores disponíveis no sistema do SISCAN, porém, no presente documento, não foram utilizadas.
  - *Rejeitada id. lâmin:* lâminas rejeitadas por erro/ausência de identificação;
  - *Rej. lâmina danif:* lâminas rejeitadas por ausência ou danificação;
  - *Rej. causas alheias:* lâminas rejeitadas por causas alheias ao laboratório;
  - *Rejeitada: Outros:* lâminas rejeitadas por outros motivos;
  - *Ins. mat. acelular:* exame com leitura prejudicada devido a material acelular ou hipocelular;

- *Ins. pres. piócitos:* exame com leitura prejudicada devido a presença de piócitos;
- *Ins. pres. art. desec.:* exame com leitura prejudicada devido a presença de artefatos de dessecamento;
- *Ins. pres. cont. exte:* exame com leitura prejudicada devido a presença de contaminantes externos;
- *Ins. pres. sup. celul:* exame com leitura prejudicada devido a intensa superposição celular;
- *Ins. pres. outros:* exame com leitura prejudicada devido a outras causas.

2.6 Trabalhos relacionados

Este capítulo descreve trabalhos recentes que estão relacionados à utilização de dados do SISCAN/DATASUS.

Em (CHAVES et al., 2022), foi analisado o impacto do rastreamento do câncer do colo do uterino no estado de Goiás durante a pandemia da Covid-19. Foram comparados os dados entre o início de 2019 ao início de 2020, obtidos a partir da base do DATASUS e concluiu-se que houve uma diminuição no número de consultas rotineiras e realização de exames no ano de 2020 em relação à 2019.

Em (GOMIDES, 2022), foi realizada uma análise sobre o impacto da pandemia da Covid-19 no rastreamento do câncer do colo do útero no município de Ouro Preto, MG. Utilizaram-se os dados do DATASUS, onde analisaram-se exames citopatológicos do colo uterino nos anos de 2019 e 2020 e observou-se uma queda de aproximadamente 60% no número de exames e indicou-se a necessidade de ações focadas para a redução na incidência do câncer do colo do útero.

Em (NEVES; EUSTÁQUIO; ARAÚJO, 2022), os autores analisaram o impacto da Covid-19 tanto no diagnóstico do câncer de colo uterino quanto no câncer de mama, avaliando dados obtidos pelo DATASUS de 2017 até 2020. A partir dos resultados obtidos, foi possível observar um aumento no número de diagnósticos de câncer de mama e câncer de colo uterino em 2017, 2018 e 2019. Porém, após o início da pandemia da Covid-19, notou-se um declínio na quantidade de diagnósticos dos elementos em estudo.

3 Metodologia

Neste capítulo, é detalhado todo o processo para elaboração dos elementos para visualização de dados temporais, sendo iniciada pela coleta e extração dos dados a partir do repositório de dados do SISCAN/DATASUS, realizando a leitura, processamento e transformação dos dados com JavaScript e Node.JS, armazenando as informações com SQL e PostgreSQL, efetuando consultas que geram valores aos dados e transformando-os em gráficos por meio do Python em conjunto com as bibliotecas Pandas (MCKINNEY et al., 2011) e Seaborn (WASKOM, 2021b). A Figura 3.1 ilustra o fluxo descrito anteriormente.

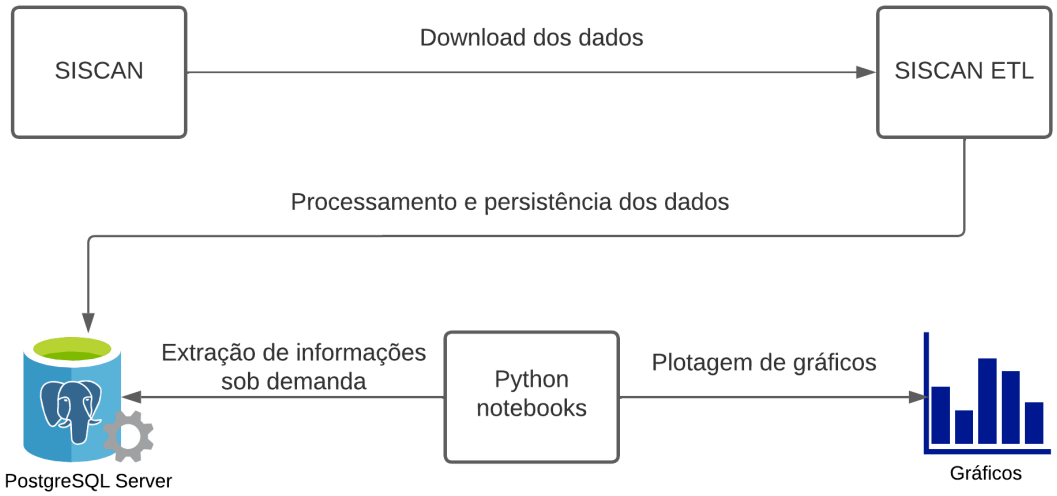


Figura 3.1 – Fluxo do sistema para visualização de dados construída no presente trabalho.

3.1 Extração dos dados no SISCAN/DATASUS

No presente trabalho, os dados foram obtidos a partir do sistema do SISCAN/DATASUS e, para cada conjunto de dados, teve-se como resultado uma matriz de duas dimensões (linha e coluna). Neste sistema, é possível selecionar o que será responsável por representar as linhas e colunas (UF, Município, Ano, Mês/Ano, Sexo e Faixa etária, por exemplo), qual será o tipo de resultado a ser avaliado (quantidade de exames ou exames alterados, por exemplo), quais serão os anos avaliados (com seleção disponível entre 2013 até 2022) e, por fim, quais serão os filtros a serem aplicados na consulta (quais serão os UFs de residência, municípios ou faixas etárias consideradas, por exemplo).

A Figura 3.2 mostra uma simulação de consulta onde serão demonstrados a UF de residência nas linhas, o ano de competência nas colunas, a quantidade de exames como medida

de avaliação, os anos de 2021 e 2022 como períodos de avaliação e os estados Acre e Alagoas como UFs disponíveis a partir do filtro de UF de residência.

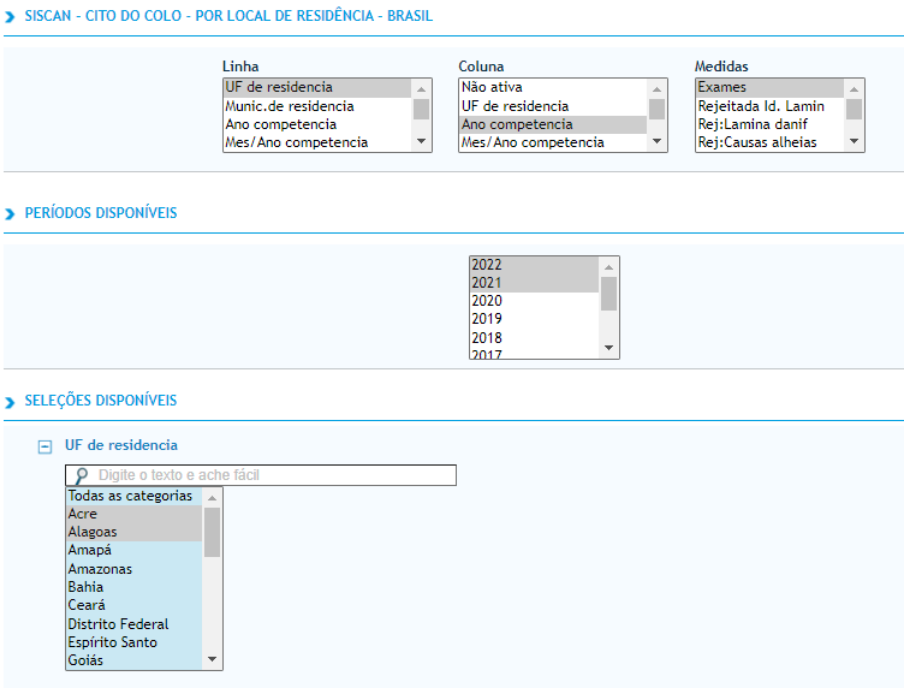
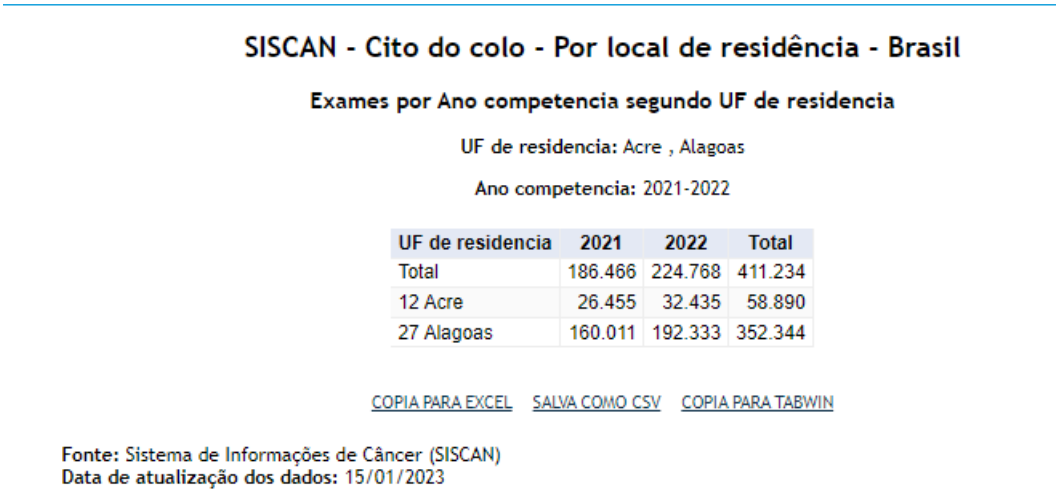


Figura 3.2 – Exemplo de consulta realizada no sistema do DATASUS. Acessado em 26/02/2023.

Para exibição dos resultados, existe a possibilidade de gerar diversos tipos de gráficos baseando-se no conjunto de dados resposta. Após solicitar a geração dos dados, o sistema exibe as informações solicitadas e disponibiliza ao usuário a opção de exportar tais informações em três formatos diferentes: Excel, CSV ou Tabwin. Para o presente trabalho, utilizou-se o csv como formato padrão para extração das informações. A Figura 3.3 expõe o resultado final tendo como base a consulta descrita anteriormente, presente na Figura 3.2.

Como este trabalho tem como objetivo avaliar todas as informações entre os anos de 2013 a 2022, separadas entre todos estados e regiões do Brasil, todas as consultas foram realizadas selecionando os mês/ano como referencial para as linhas, a UF de residência como referencial para as colunas, todos os anos entre 2013 e 2022 como período de avaliação, nenhuma aplicação de filtro e a realização de uma consulta por vez para obter as informações sobre cada indicador selecionado, tendo como base a lista de indicadores informados na Seção 2.5. Para realização deste trabalho, a coleta e extração dos dados foi realizada de maneira manual e não foram encontradas limitações para realizações de consultas no sistema.



SISCAN - Cito do colo - Por local de residência - Brasil

Exames por Ano competencia segundo UF de residencia

UF de residencia: Acre , Alagoas

Ano competencia: 2021-2022

UF de residencia	2021	2022	Total
Total	186.466	224.768	411.234
12 Acre	26.455	32.435	58.890
27 Alagoas	160.011	192.333	352.344

COPIA PARA EXCEL SALVA COMO CSV COPIA PARA TABWIN

Fonte: Sistema de Informações de Câncer (SISCAN)
Data de atualização dos dados: 15/01/2023

Figura 3.3 – Exemplo de resultado de consulta realizada no sistema do DATASUS. Acessado em 26/02/2023.

3.2 Processamento e persistência dos dados

Para utilização dos conjuntos de dados obtidos na seção anterior, foi necessário realizar um processo de normalização e padronização dos dados, o que, em conjunto com o processo de extração dos dados, pode ser definido como um processo de ETL (*Extract, Transform and Load*). O processo de transformação dos dados foi implementado utilizando a linguagem de programação Javascript por meio da plataforma de desenvolvimento Node.JS e a persistência dos dados tratados foi implementada utilizando o Sistema de Gerenciamento de Banco de Dados Relacional PostgreSQL. A seguir, será demonstrado como o processo foi montado para promover a persistência dos dados com sucesso.

3.2.1 Persistência dos dados

Para que as informações sejam processadas de maneira uniforme, permitindo agrupamentos entre diversos tipos de medidas, regiões, estados e períodos, criou-se uma tabela no banco de dados denominada de *exams*, demonstrada na Figura 3.4.

Cada registro dessa tabela tem a responsabilidade de informar sobre uma única linha em relação as consultas realizadas no processo de extração dos dados, informando sobre o ano (coluna *year*) e mês de referência (coluna *month*), estado (coluna *state*), região daquele estado (coluna *region*), tipo de medida (coluna *exam_type*) e sua quantidade observada (coluna *amount*). Foram incluídas duas outras informações na tabela (identificador pela coluna *id* e data de criação pela coluna *created_at*), mas ambas tem objetivo somente técnico e não tem qualquer tipo de relação com o material de estudo.

Todas as coluna dessa tabela são denominadas de atributos simples, monovalorados e

descritivos. Além disso, não apresenta qualquer tipo de relacionamento com outra tabela no banco de dados.

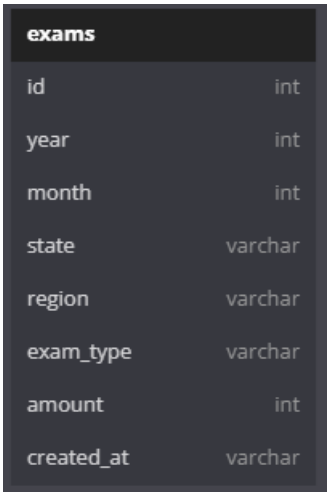


Figura 3.4 – Diagrama de entidade-relacionamento utilizado no projeto.

3.2.2 Processamento dos dados

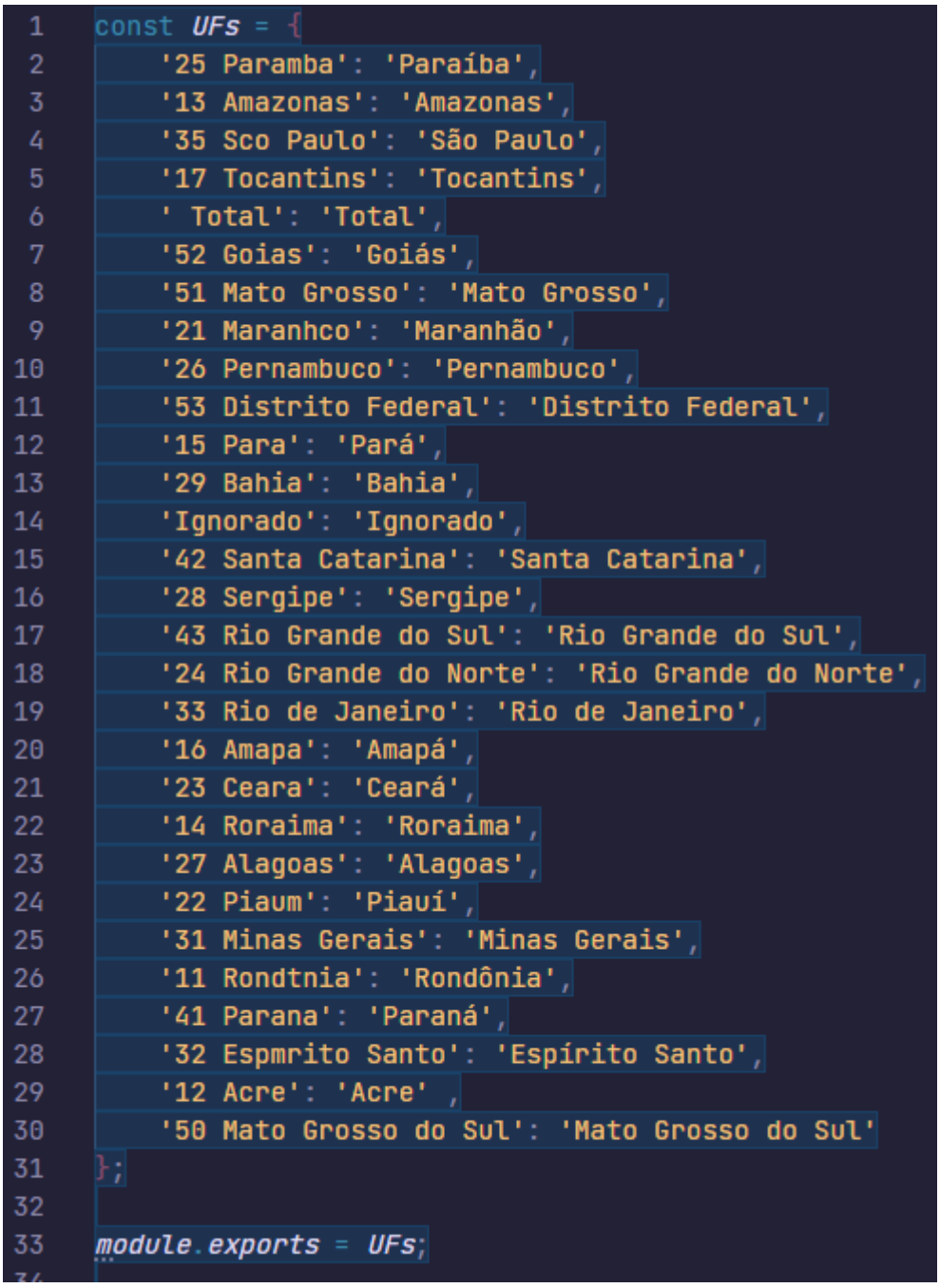
Para ter um conjunto uniforme e sanitizado de todas as consultas realizadas no DATASUS, além da possibilidade de cruzar diferentes indicadores e gerar *insights* a partir de tais dados, foi necessário efetuar o seu processamento. Para isso, as seguintes etapas abaixo foram realizadas.

3.2.2.1 Leitura dos arquivos

O sistema tem como objetivo ler todos os arquivos com formato csv presentes em uma pasta no repositório do projeto. Cada arquivo tem como seu nome o tipo de medida presente em seus resultados. Com isso, o algoritmo lê linha a linha de todos os arquivos, com exceção da primeira (linha de cabeçalho). A primeira coluna fica responsável por indicar qual é o estado de referência e todas as colunas subsequentes são responsáveis por demonstrar as suas respectivas quantidades do ano de referência.

3.2.2.2 Tratamento de informação: UF de residência

Notou-se a partir da extração das informações (exemplo na Figura 3.3) que as informações referentes à UF de residência possuem ruídos. Além de existirem números como prefixo do estado sem quaisquer motivo aparente, notou-se também a presença de uma decodificação que não permite a inclusão de acentos. Com isso, foi necessário criar um dicionário para cada UF encontrada e, assim, realizar sua substituição com as informações sem ruídos. A Figura 3.5 mostra o dicionário criado para mapeamento da substituição das informações.



```
const UFs = {
    '25 Paramba': 'Paraíba',
    '13 Amazonas': 'Amazonas',
    '35 Sco Paulo': 'São Paulo',
    '17 Tocantins': 'Tocantins',
    ' Total': 'Total',
    '52 Goias': 'Goiás',
    '51 Mato Grosso': 'Mato Grosso',
    '21 Maranhco': 'Maranhão',
    '26 Pernambuco': 'Pernambuco',
    '53 Distrito Federal': 'Distrito Federal',
    '15 Para': 'Pará',
    '29 Bahia': 'Bahia',
    'Ignorado': 'Ignorado',
    '42 Santa Catarina': 'Santa Catarina',
    '28 Sergipe': 'Sergipe',
    '43 Rio Grande do Sul': 'Rio Grande do Sul',
    '24 Rio Grande do Norte': 'Rio Grande do Norte',
    '33 Rio de Janeiro': 'Rio de Janeiro',
    '16 Amapa': 'Amapá',
    '23 Ceara': 'Ceará',
    '14 Roraima': 'Roraima',
    '27 Alagoas': 'Alagoas',
    '22 Piaum': 'Piauí',
    '31 Minas Gerais': 'Minas Gerais',
    '11 Rondtnia': 'Rondônia',
    '41 Parana': 'Paraná',
    '32 Espmrito Santo': 'Espírito Santo',
    '12 Acre': 'Acre' ,
    '50 Mato Grosso do Sul': 'Mato Grosso do Sul'
};

module.exports = UFs;
```

Figura 3.5 – Dicionário para tratamento das UFs de residência.

3.2.2.3 Processamento de nova informação: região

Como a informação referente a região do Brasil no qual tal estado se encontra não está disponível nas informações do DATASUS, foi necessário fazer um processamento de forma a criar tal informação e inseri-la no banco de dados. Para tal, desenvolveu-se também um dicionário responsável por mapear em qual região aquele estado está incluso. A Figura 3.6 mostra o dicionário utilizado para criar tal informação.

```
const regions = {
    'Rondônia':             'Norte',
    'Acre':                 'Norte',
    'Amazonas':             'Norte',
    'Roraima':              'Norte',
    'Pará':                 'Norte',
    'Tocantins':            'Norte',
    'Amapá':                'Norte',
    'Maranhão':             'Nordeste',
    'Piauí':                'Nordeste',
    'Ceará':                'Nordeste',
    'Rio Grande do Norte':  'Nordeste',
    'Paraíba':              'Nordeste',
    'Pernambuco':           'Nordeste',
    'Alagoas':              'Nordeste',
    'Sergipe':              'Nordeste',
    'Bahia':                'Nordeste',
    'Minas Gerais':         'Sudeste',
    'Espírito Santo':       'Sudeste',
    'Rio de Janeiro':       'Sudeste',
    'São Paulo':            'Sudeste',
    'Paraná':               'Sul',
    'Santa Catarina':       'Sul',
    'Rio Grande do Sul':    'Sul',
    'Mato Grosso do Sul':   'Centro-Oeste',
    'Mato Grosso':          'Centro-Oeste',
    'Goiás':                'Centro-Oeste',
    'Distrito Federal':     'Centro-Oeste',
    'Ignorado':             'Not a region',
    ' Total':               'Not a region',
    'Mes/Ano competencia':  'Not a region'
};

module.exports = regions;
```

Figura 3.6 – Dicionário para transformação das UFs de residência em regiões do Brasil.

3.2.2.4 Persistência das informações no banco de dados

Após o processamento e criação de todas as informações necessárias, gerou-se diversos comandos SQL responsáveis por inserir os dados em um banco de dados alocado em ambiente local. Para cada linha analisada, criam-se comandos SQL capazes de inserir cada célula da matriz contendo a quantidade responsável por indicar o resultado referente ao indicador. A Figura 3.7 apresenta um exemplo de geração de um comando SQL a partir dos resultados presentes em uma linha-resultado. Posteriormente, o sistema é responsável por estabelecer uma conexão com o banco de dados e realizar a inserção de cada informação gerada.

UF de residencia	NOVEMBRO/2022	DEZEMBRO/2022	Total
Total	4.355	2.693	7.048
12 Acre	4.355	2.693	7.048

```sql
INSERT INTO exams (year, month, state, region, exam_type, amount)
    VALUES(2022, 11, 'Total', 'Not a region', 'Exames', 4355);
INSERT INTO exams (year, month, state, region, exam_type, amount)
    VALUES(2022, 12, 'Total', 'Not a region', 'Exames', 2693);
INSERT INTO exams (year, month, state, region, exam_type, amount)
    VALUES(2022, 11, 'Acre', 'Norte','Exames', 4355);
INSERT INTO exams (year, month, state, region, exam_type, amount)
    VALUES(2022, 12, 'Acre', 'Norte','Exames', 2693);
```

Figura 3.7 – Comandos de INSERT em SQL a partir de uma matriz de resultado.

3.3 Extração de informações sob demanda

Com o armazenamento uniforme de todas as informações, teve-se como próximo objetivo a criação de consultas sob demanda que aplicam determinado tipo de inteligência acima do conjunto de dados obtido, seja agregando informações separando dados por região, por estado, por indicador ou grupos de indicadores. Para isso, utilizou-se da linguagem SQL para criação de consultas no conjunto de dados, onde cada consulta teria responsabilidade de retornar as informações necessárias sobre somente uma análise. A Figura 3.8 demonstra um exemplo utilizando o comando *select* do SQL, responsável por buscar informações presentes no banco de dados. A Tabela 3.1 mostra o resultado obtido a partir da consulta realizada.

```sql
select
    year,
    sum(amount) as quantidade
from exams e
where region in ('Norte')
    and exam_type in ('Exames')
group by year
order by year asc
;
```

Figura 3.8 – Comando *select* do SQL com o objetivo de obter a quantidade de exames por ano da região Norte.

	123 year	123 quantidade
1	2,013	8,863
2	2,014	118,449
3	2,015	179,241
4	2,016	257,522
5	2,017	307,845
6	2,018	441,114
7	2,019	479,439
8	2,020	274,584
9	2,021	468,528
10	2,022	610,727

Tabela 3.1 – Quantidade de exames de Papanicolau realizados na região Norte, Brasil, de 2013 a 2022.

3.4 Predição de valores durante a pandemia

Considerando os dados temporais do SISCAN anteriores ao primeiro ano da pandemia (2020), foi elaborada uma estratégia para estimar os valores dos dados em consideração a partir de dois passos. Salienta-se que os valores estimados, ou seja simulados, são úteis para comparar com os dados reais coletados durante a pandemia da Covid-19, permitindo avaliar o seu impacto no número de exames de Papanicolau.

No passo 1 da estratégia, obtém-se a taxa de crescimento médio considerando os dados consolidados dos anos de 2017, 2018 e 2019 através da equação:

$$\textbf{taxa\_crescimento} = \frac{\frac{valor\_consolidado\_2019}{valor\_consolidado\_2018} + \frac{valor\_consolidado\_2018}{valor\_consolidado\_2017}}{2}. \tag{3.1}$$

No passo 2, calculam-se os resultados das estimativas (valores simulados) dos indicadores em determinado ano da seguinte forma:

$$\textbf{valor\_simulado\_2020} = valor\_consolidado\_2019 * taxa\_crescimento, \tag{3.2}$$

$$\textbf{valor\_simulado\_2021} = valor\_simulado\_2020 * taxa\_crescimento, \tag{3.3}$$

$$\textbf{valor\_simulado\_2022} = valor\_simulado\_2021 * taxa\_crescimento. \tag{3.4}$$

3.5 Elaboração dos gráficos

Com as consultas elaboradas e os valores das simulações calculadas, o último processo ficou responsável pela elaboração de visualizações gráficas. Para tal, definiu-se que seriam gerados grupos de gráficos, onde cada grupo seria responsável por demonstrar um tipo de informação. Os seguintes grupos foram elaborados:

- Grupos representando cada tipo de resultado do exame (quantidade de exames, resultados negativos, resultados alterados, resultados alterado escamoso e resultados alterado glandular), onde cada gráfico do grupo representa uma região do Brasil;
- Grupos representando cada tipo de resultado do exame (quantidade de exames, resultados negativos, resultados alterados, resultados alterado escamoso e resultados alterado glandular) e separando cada estado da região em uma série no gráfico, onde cada gráfico do grupo representa uma região do Brasil;
- Grupo representando todos os resultados de exames, onde cada gráfico do grupo representa uma região do Brasil;
- Grupo representando os resultados dos exames relacionados a resultados alterados, onde cada gráfico do grupo representa uma região do Brasil.

No Capítulo 4 são apresentados os grupos de gráficos produzidos conforme detalhados anteriormente.

4 Resultados Obtidos

O presente capítulo apresenta os resultados alcançados, dentro do tema proposto, a partir dos dados obtidos do SISCAN/DATASUS. Os dados representam informações entre os anos de 2013 a 2022. A partir do primeiro ano de pandemia, foi feita uma simulação de valores para os próximos anos, respeitando uma média de taxa de crescimento obtida por meio das taxas de crescimento dos três anos que antecedem a Covid-19. O processo de consulta aos dados foi realizado por meio de consultas SQL e os gráficos para análise foram elaborados utilizando as bibliotecas Pandas e Seaborn.

4.1 Quantidade de exames por região do Brasil

Os resultados das análises aqui apresentados são referentes a quantidade de exames realizados em todo o Brasil entre o período de 2013 a 2022. Cada gráfico da Figura 4.1 representa uma região do Brasil e, em cada uma deles, a linha contínua em azul representa os dados obtidos enquanto a linha amarela tracejada representa a simulação dos dados obedecendo a taxa de crescimento entre os anos de 2017, 2018 e 2019.

As regiões Nordeste e Sudeste apresentam praticamente as mesmas informações quantitativas, ambas consideradas como regiões com mais exames de Papanicolau realizados no Brasil. Porém, a região Nordeste se destaca pelos dados simulados serem consideravelmente maiores do que foi registrado no ano de 2022. Para a região Sudeste a simulação praticamente representa a realidade no último ano da análise. Em 2022, de acordo com a Tabela 4.1, a região Nordeste deveria apresentar um crescimento de aproximadamente 21% em relação a 2019 e apresentou um crescimento de apenas 4%, enquanto a região Sudeste deveria apresentar um crescimento de 6% e cresceu 4%.

As regiões Norte e Centro-Oeste apresentaram as menores taxas de exames realizados no período. Apesar de não alcançarem as quantidades de exames definidas pela simulação, ambas apresentaram retomadas suficientemente boas em 2022 de forma a superarem a quantidade de exames realizados no último ano anterior à pandemia.

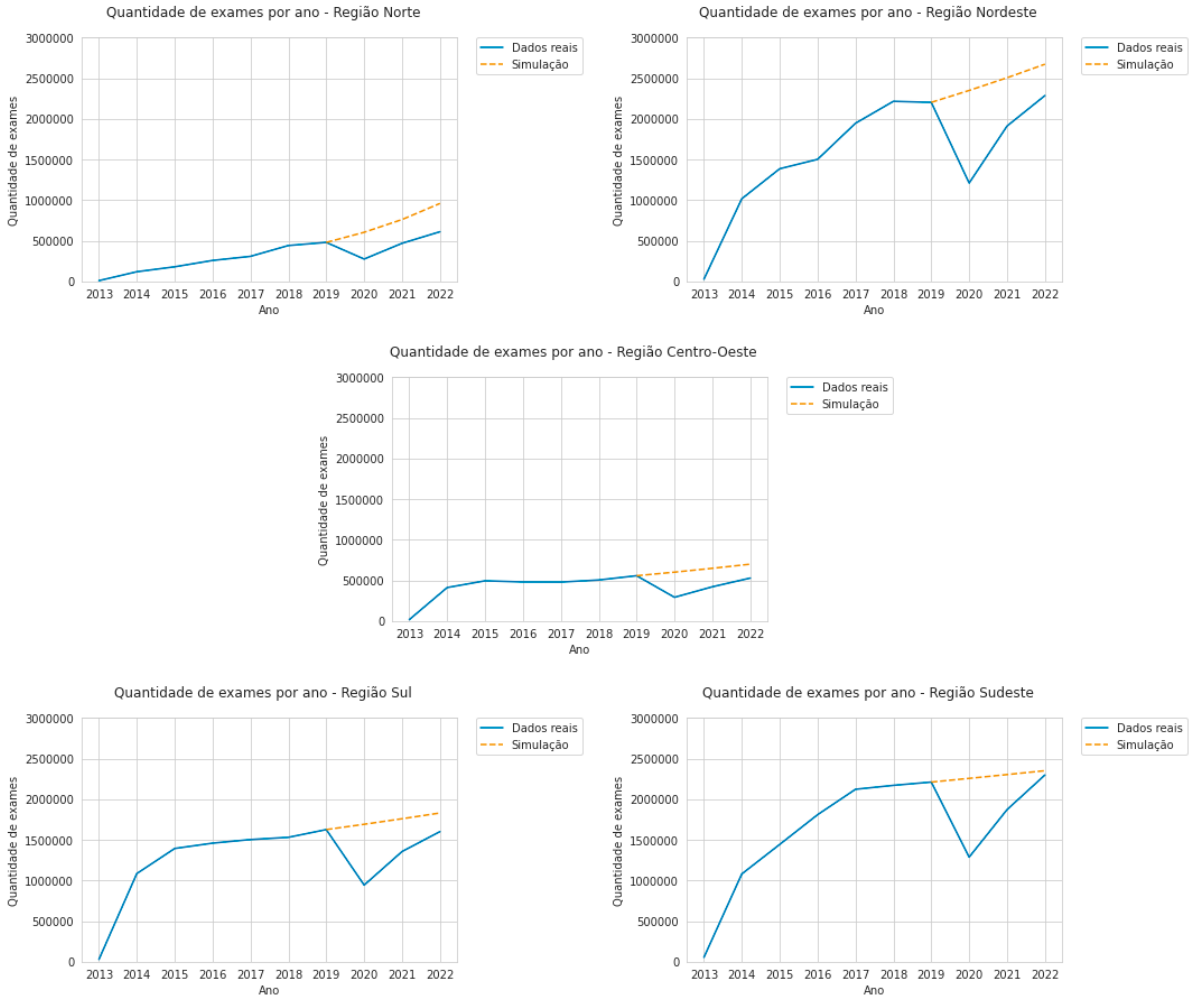


Figura 4.1 – Quantidade de exames por ano em cada uma das regiões do Brasil.

	2019	2020		2021		2022	
		Real	Simulação	Real	Simulação	Real	Simulação
Norte	-	-43%	26%	-2%	59%	27%	100%
Nordeste	-	-45%	7%	-13%	14%	4%	21%
Centro-Oeste	-	-48%	8%	-24%	16%	-5%	26%
Sul	-	-42%	4%	-17%	8%	-2%	13%
Sudeste	-	-42%	2%	-15%	4%	4%	6%

Tabela 4.1 – Porcentagens de crescimento real e projetados sobre a quantidade de exames por ano em cada uma das regiões do Brasil.

4.2 Resultados negativos por região do Brasil

Os resultados apresentados a seguir são referentes aos resultados negativos encontrados, presentes na Figura 4.2, em cada região do Brasil no período definido pelo documento.

Neste caso, os resultados apresentaram o mesmo padrão identificado naqueles apresentados na seção de quantidade de exames (Seção 4.1). As regiões Nordeste e Sudeste apresentaram os maiores valores quantitativos. As regiões Norte e Centro-oeste com os menores valores e, por fim, os valores simulados sempre se apresentaram superiores em comparação aos dados reais.

Analisando a Tabela 4.2, notou-se que a maior queda real foi encontrada na região Centro-oeste em 2020, com -48%, enquanto a menor queda simulada ocorre nas regiões Sul/Sudeste com 4% de crescimento. Em relação ao maior crescimento real identificado, teve-se a região Norte com 29% em 2022, enquanto o maior crescimento simulado ocorreu também na mesma região e mesmo período, apresentando 95% de crescimento.

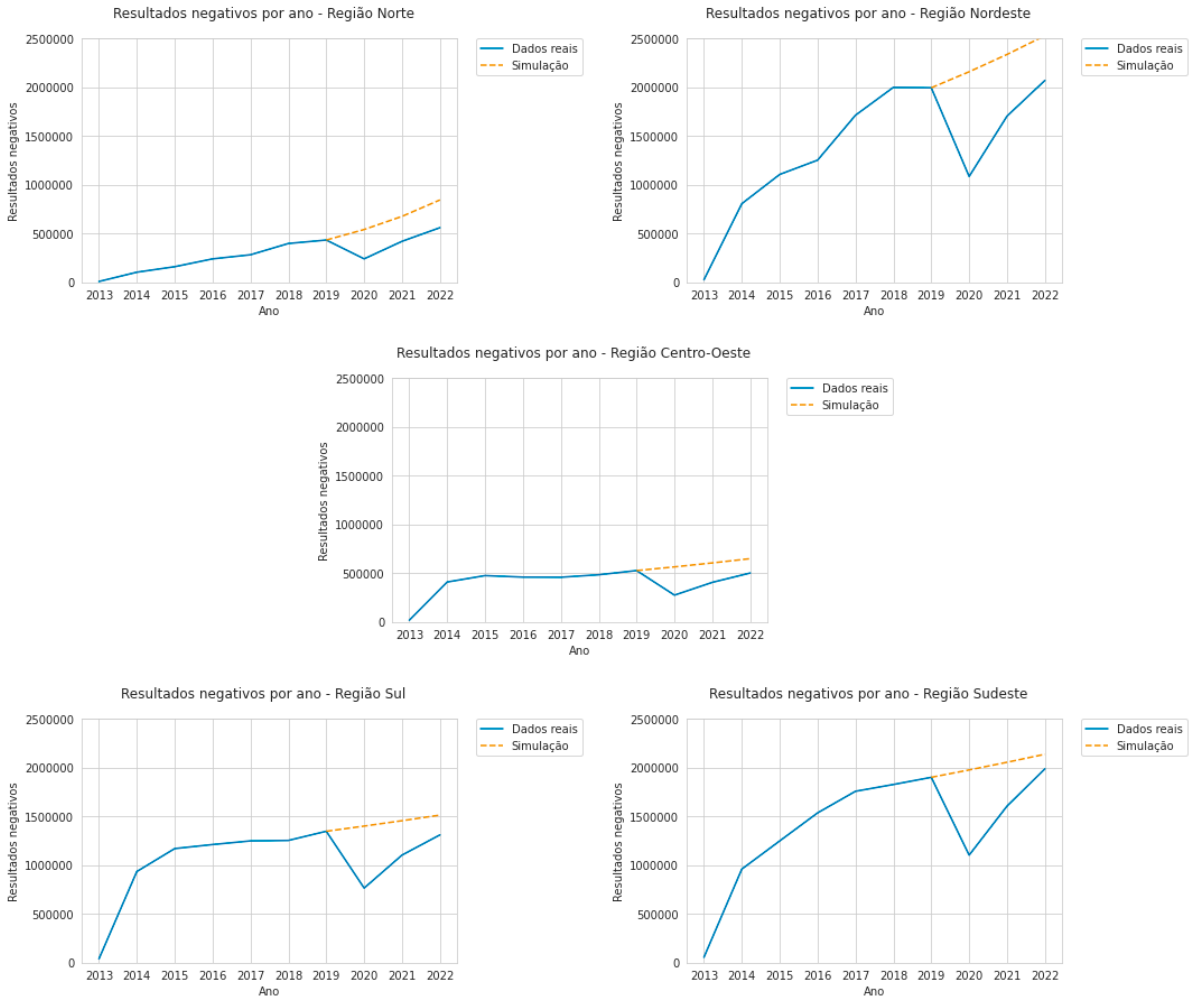


Figura 4.2 – Resultados negativos por ano em cada uma das regiões do Brasil.

	2019	2020		2021		2022	
		Real	Simulação	Real	Simulação	Real	Simulação
Norte	-	-44%	25%	-3%	56%	29%	95%
Nordeste	-	-46%	8%	-15%	17%	4%	27%
Centro-Oeste	-	-48%	7%	-23%	15%	-5%	23%
Sul	-	-43%	4%	-18%	8%	-3%	12%
Sudeste	-	-42%	4%	-16%	8%	5%	12%

Tabela 4.2 – Porcentagens de crescimento real e projetados sobre os resultados negativos por ano em cada uma das regiões do Brasil.

4.3 Resultados alterados por região do Brasil

Nesta seção, são apresentados na Figura 4.3 os resultados referentes aos resultados alterados nas regiões do Brasil. Neste caso, as regiões Nordeste, Sul e Sudeste apresentaram praticamente os mesmos resultados, com a diferença de que a região Sul apresentou os maiores valores simulados dentre os citados. A região Centro-Oeste apresentou os menores valores quantitativos. Em nenhuma região os valores reais chegaram a apresentar proximidade ao resultado simulado.

De acordo com a Tabela 4.3, a maior queda real foi identificada na região Centro-Oeste, com -44% em 2020 e o menor crescimento simulado ocorreu também em 2020, na região Sudeste, com 8%. Os maiores crescimentos reais e simulados foram identificados em 2021 e 2022 na região Norte, com 96% e 174%, respectivamente.

	2019	2020		2021		2022	
		Real	Simulação	Real	Simulação	Real	Simulação
Norte	-	-33%	40%	7%	96%	31%	174%
Nordeste	-	-40%	15%	-2%	33%	21%	53%
Centro-Oeste	-	-44%	18%	-15%	38%	20%	63%
Sul	-	-35%	17%	-12%	37%	0%	60%
Sudeste	-	-32%	8%	-8%	16%	3%	25%

Tabela 4.3 – Porcentagens de crescimento real e projetados sobre os resultados alterados por ano em cada uma das regiões do Brasil.

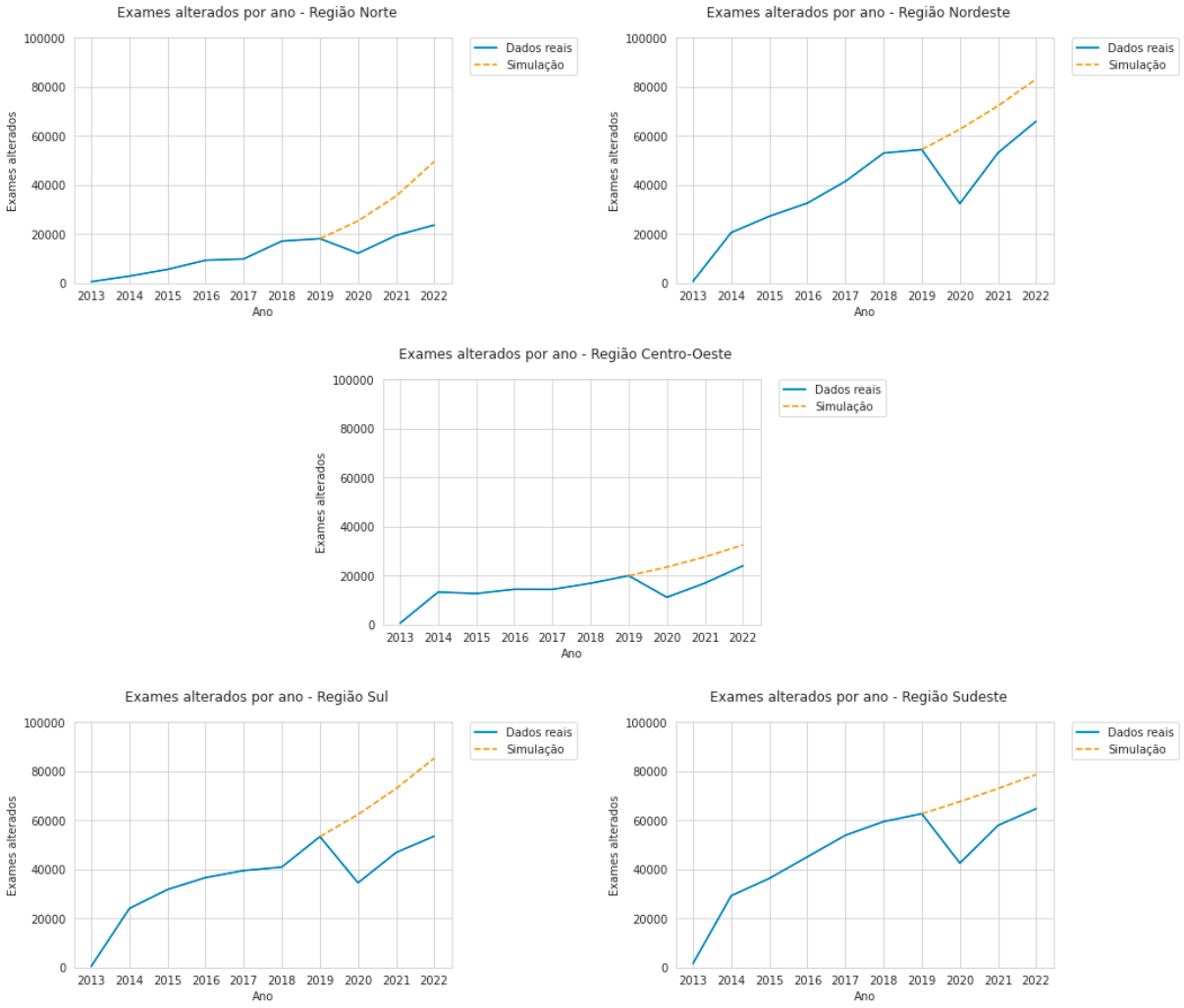


Figura 4.3 – Resultados alterados por ano em cada uma das regiões do Brasil.

4.4 Resultados alterado escamoso por região do Brasil

De acordo com a Figura 4.4, as regiões Nordeste e Sudeste apresentaram praticamente os mesmos resultados reais, com uma queda mais significativa para a região Sudeste no ano de 2020. Em relação aos dados simulados, as regiões Nordeste e Sul apresentaram os maiores resultados, indicando que ambas apresentaram as mesmas taxas de crescimento nos últimos 3 anos anteriores a pandemia. As regiões Norte e Centro-Oeste apresentaram os menores resultados apresentados, seja em dados reais ou simulados.

De acordo com a Tabela 4.4, a maior queda real foi identificada na região Centro-Oeste, com -44% em 2020 e o menor crescimento simulado ocorreu também em 2020, na região Sudeste, com 8%. Os maiores crescimentos reais e simulados foram identificados em 2021 e 2022 na região Norte, com 95% e 173%, respectivamente.

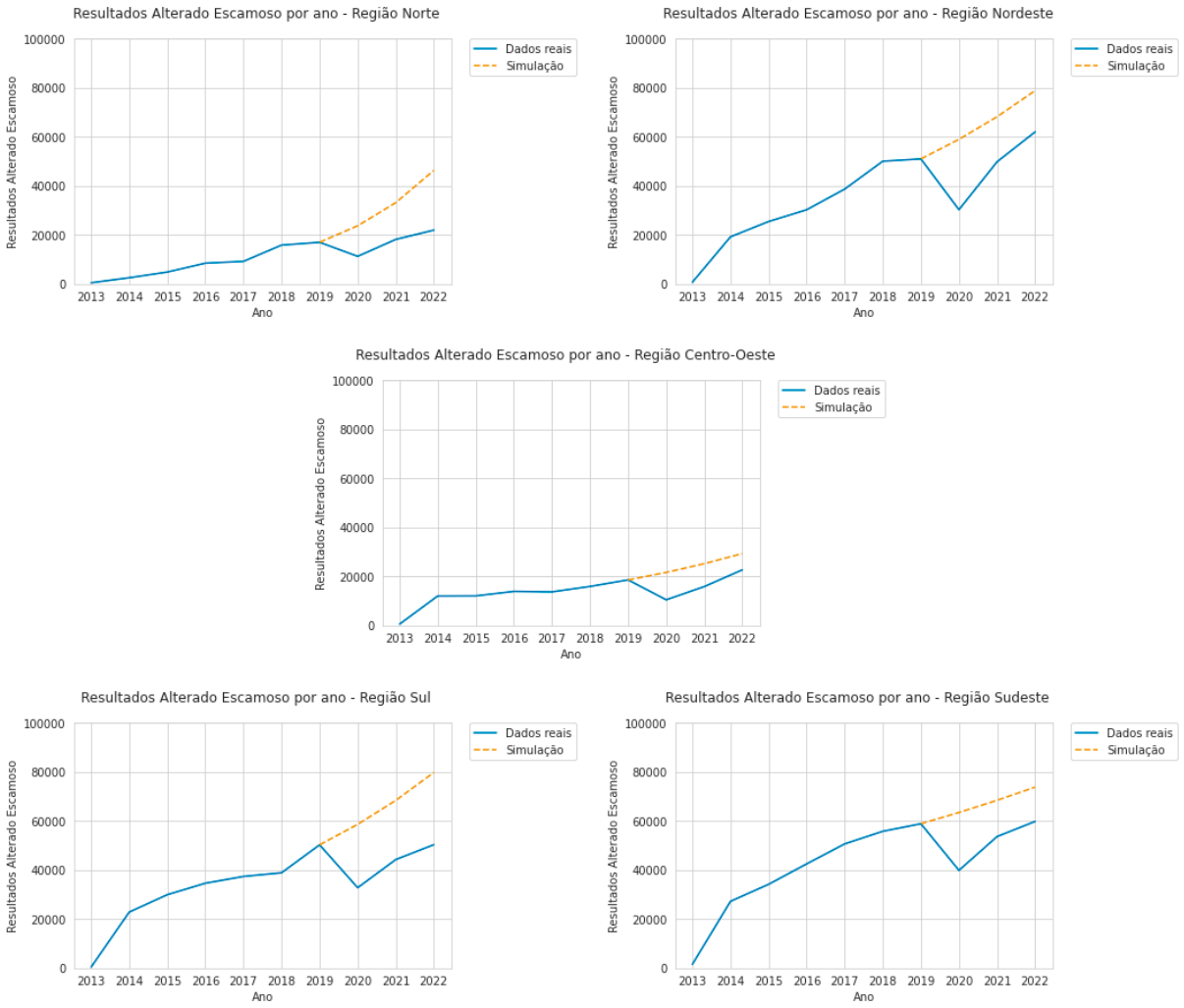


Figura 4.4 – Resultados alterados escamoso por ano em cada uma das regiões do Brasil.

	2019	2020		2021		2022	
		Real	Simulação	Real	Simulação	Real	Simulação
Norte	-	-34%	40%	7%	95%	29%	173%
Nordeste	-	-41%	16%	-2%	34%	22%	55%
Centro-Oeste	-	-44%	16%	-15%	36%	22%	58%
Sul	-	-35%	17%	-12%	36%	0%	59%
Sudeste	-	-32%	8%	-9%	16%	2%	25%

Tabela 4.4 – Porcentagens de crescimento real e projetados sobre os resultados alterado escamoso por ano em cada uma das regiões do Brasil.

4.5 Resultados alterado glandular por região do Brasil

Foi identificado, conforme mostrado na Figura 4.5, que a região Sudeste apresentou os maiores resultados alterados glandulares e, curiosamente, foi o único caso onde os dados reais da competência de 2022 superou os dados simulados do mesmo período. A região Sul apresentou os maiores dados simulados e, consequentemente, a maior taxa de crescimento. A região Centro-Oeste apresentou os menores resultados reais e simulados.

A Tabela 4.5 apresenta as taxas de crescimento real e simulada tendo como ano base 2019. A partir disso, observou-se que as menores taxas de crescimento ocorreram em 2020 nas regiões Centro-Oeste e Sul com -44% e -43%, respectivamente. Em 2022, observou-se a maior taxa de crescimento real e simulada, ambas na região Norte, com 55% de crescimento real e 232% no crescimento simulado.

	2019	2020		2021		2022	
		Real	Simulação	Real	Simulação	Real	Simulação
Norte	-	-19%	49%	14%	122%	55%	232%
Nordeste	-	-37%	12%	-3%	26%	17%	41%
Centro-Oeste	-	-44%	36%	-13%	85%	6%	151%
Sul	-	-43%	29%	-15%	67%	7%	117%
Sudeste	-	-31%	8%	8%	18%	29%	28%

Tabela 4.5 – Porcentagens de crescimento real e projetados sobre sobre os resultados alterado glandular por ano em cada uma das regiões do Brasil.

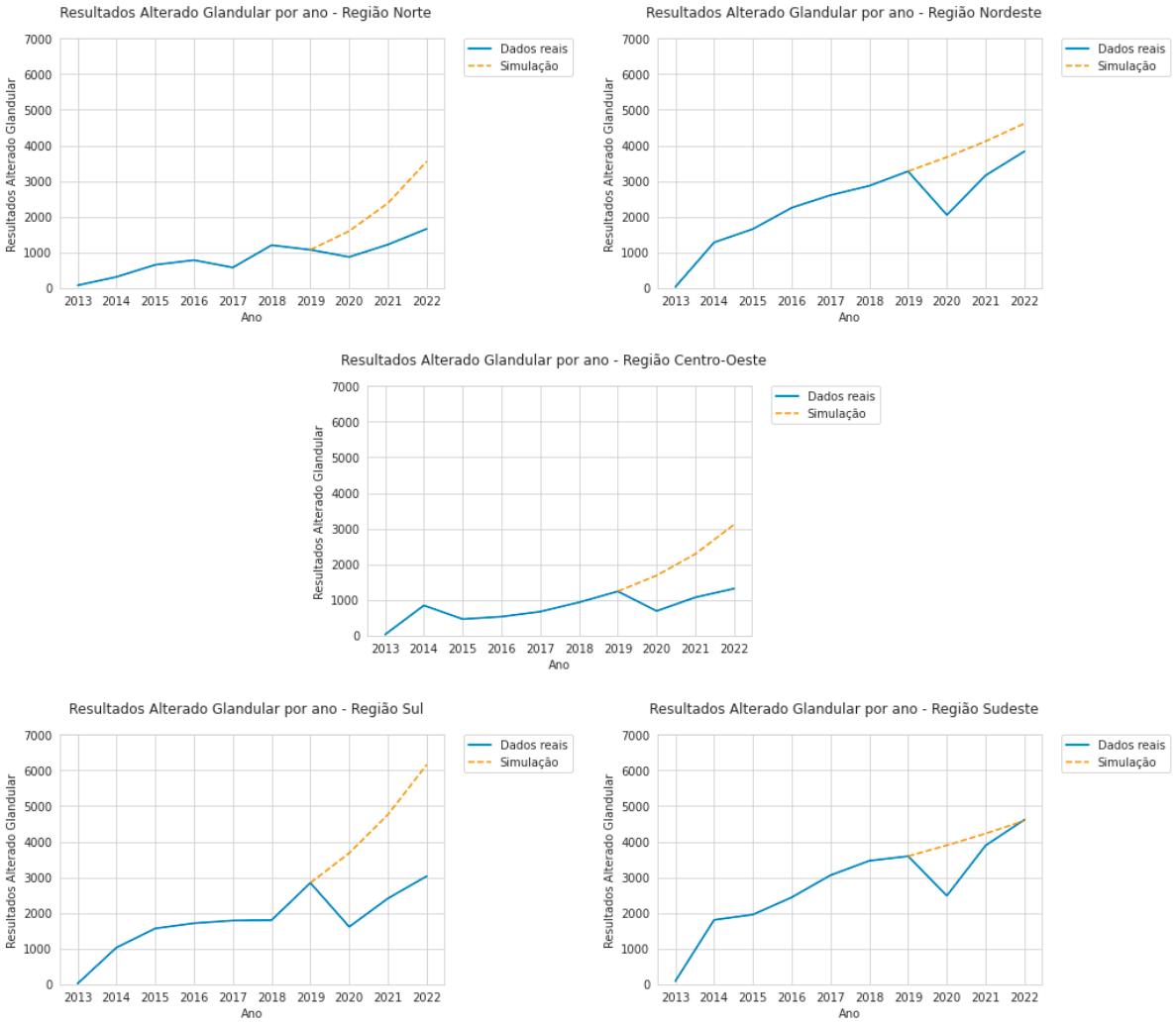


Figura 4.5 – Resultados alterado glandular por ano em cada uma das regiões do Brasil.

4.6 Quantidade de exames por cada estado de uma região do Brasil

Os dados mostrados na Figura 4.6 são referentes a quantidade de exames realizados por ano em todo Brasil, entre o período de 2013 a 2022. Cada gráfico retrata uma região do Brasil e cada série representa as informações de determinado estado.

Na região Norte, apresenta-se o Pará como o estado com maior número de exames no decorrer dos anos e os outros estados se mantendo aproximadamente iguais, com exceção do estado do Amazonas nos anos de 2021 e 2022, onde teve-se um aumento considerável de exames identificados.

Já em relação as regiões Nordeste e Sudeste, manteve-se números de exames praticamente iguais durante o período demonstrado. Destacam-se quantitativamente os estados da Bahia e Pernambuco da região Norte e Minas Gerais e São Paulo da região Sudeste por representarem um grande número de exames realizados em comparação aos demais estados da região.

Por fim, as regiões Centro-Oeste e Sul tiveram um número de exames praticamente iguais durante todo o período para cada estado analisado, com exceção do Distrito Federal, destacado pelo grande aumento de exames a partir do ano de 2018.

De maneira geral, observou-se um padrão em todos os estados. Ocorre uma diminuição representativa na quantidade de exames no primeiro ano da pandemia (2020) e uma retomada nos próximos dois anos, onde, na maioria dos casos, constata-se a volta ou até mesmo a superação dos números presentes no último ano anterior a Covid-19.

Por fim, o estado do Rio de Janeiro foi identificado como incomum em relação ao padrão identificado, pois, curiosamente, os exames aumentaram mesmo nos anos de maior impacto da Covid-19 (2020 e 2021), chegando a ter um aumento de praticamente 100% no número de exames em 2022 levando como consideração o ano base de 2020.

Figura 4.6 – Quantidade de exames de cada estado em cada uma das regiões do Brasil.

4.7 Resultados negativos por cada estado de uma região do Brasil

Os resultados da presente seção são referentes a quantidade de exames negativos realizados por ano em todo Brasil entre o período de 2013 a 2022, presentes na Figura 4.7. Cada gráfico representa uma região do Brasil e, em cada série do gráfico, são representados os dados de cada estado.

As regiões Norte e Centro-Oeste demonstram as menores quantidades de resultados negativos dentre as regiões por consequência de serem, também, as regiões com menores quantidades de exames realizados.

As regiões Nordeste, Sul e Sudeste apresentaram o maior volume de resultados negativos do país, com destaque para a região Sudeste e os estados de Minas Gerais e São Paulo por apresentarem os maiores valores em relação aos demais estados do país.

Por fim, conforme identificado na Figura 4.6, o número de exames mostrou queda considerável no ano de início da pandemia, em 2020, e, posteriormente, decorreu de voltar as informações contidas anteriormente. Na Figura 4.7 pode ser identificado este mesmo padrão.

Figura 4.7 – Resultados negativos de cada estado em cada uma das regiões do Brasil.

4.8 Resultados alterados por cada estado de uma região do Brasil

Os resultados aqui apresentados são referentes a quantidade de resultados alterados realizados por ano em todo Brasil entre o período de 2013 a 2022. Cada gráfico da Figura 4.8 representa uma região do Brasil e, em cada série do gráfico, são representados os dados de cada estado.

A região Sudeste possui a maior porção de resultados alterados dentre as demais regiões, com ênfase nos estados de São Paulo e Minas Gerais. Os estados do Espírito Santo e Rio de Janeiro se mantém com poucos resultados alterados durante os anos em relação a seus vizinhos, porém, no estado do Rio de Janeiro, a taxa de crescimento de um ano em relação ao ano anterior foi sempre positiva, com exceção do último ano de análise, assim como o que fora identificado na Figura 4.6.

A região Norte possui o estado do Pará com maior porção de informação, assim como nas Figuras 4.6 e 4.7, por fazer mais testes em relação a outros estados. Entretanto, o estado do Amazonas apresenta um grande aumento na taxa de crescimento em relação ao ano anterior nos anos de 2021 e 2022, anos estes posteriores ao primeiro ano da pandemia.

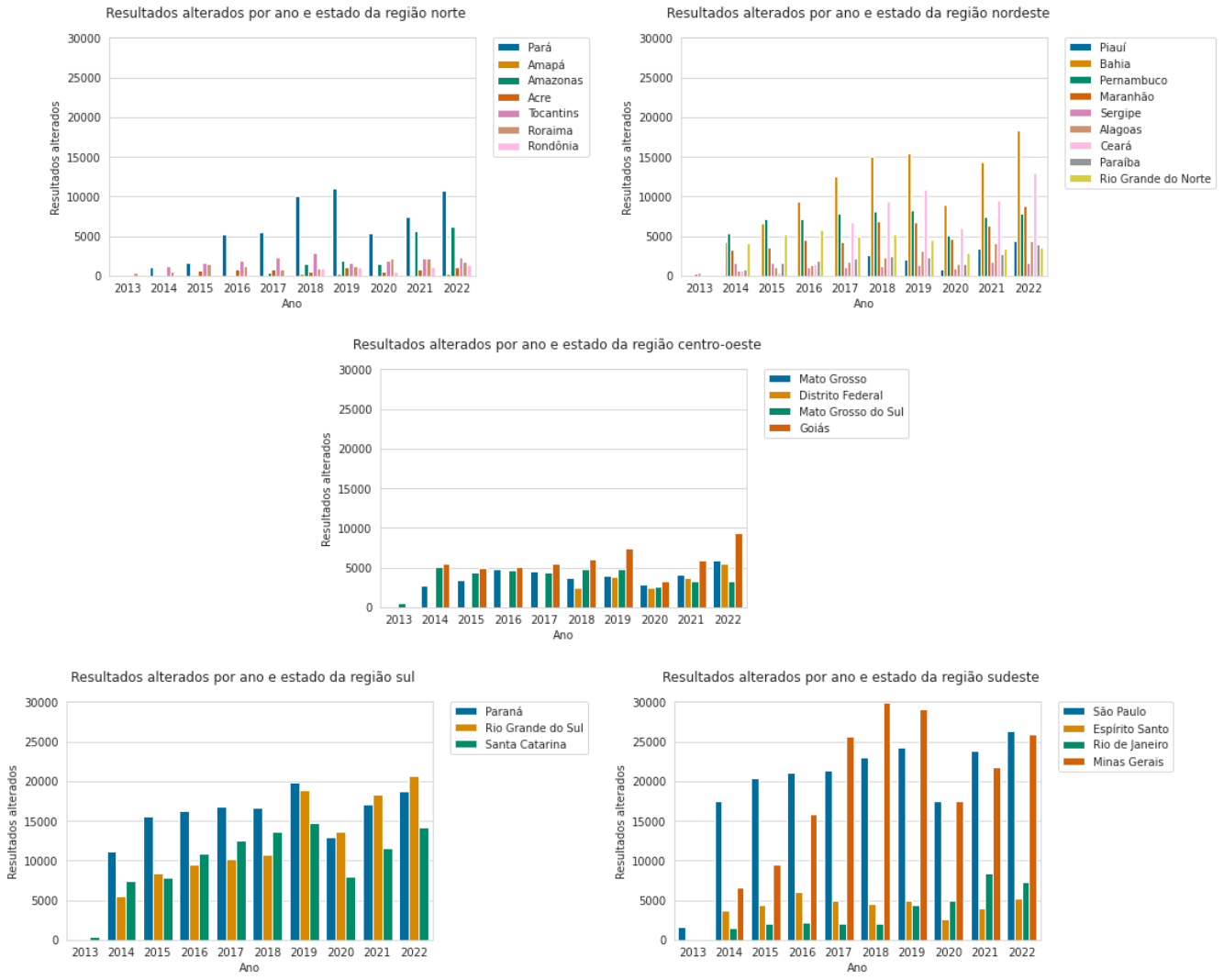


Figura 4.8 – Resultados alterados de cada estado em cada uma das regiões do Brasil.

4.9 Resultados alterado escamoso por cada estado de uma região do Brasil

Os resultados desta seção apresentam a quantidade de resultados alterado escamoso por ano em todo Brasil, entre o período de 2013 a 2022, mostrados na Figura 4.9. Cada gráfico representa uma região do Brasil e, em cada série do gráfico, são representados os dados de cada estado.

O maior número de resultados alterado escamoso estão localizados na região Sudeste e os estados de Minas Gerais e São Paulo agregam a maior porção das informações.

Em relação às regiões do Nordeste, Centro-Oeste e Sul, não foi identificado uma taxa de crescimento significante de um ano para o outro, com exceção dos estados da Bahia (2020 e 2021) e Goiás (2021 e 2022).

De maneira geral, a figura mostra praticamente os mesmos gráficos identificados na Figura 4.8, com a diferença na proporção das informações, visto que indicador de resultado alterado escamoso é um subconjunto dos resultados alterados.

Figura 4.9 – Resultados alterado escamoso de cada estado em cada uma das regiões do Brasil.

4.10 Resultados alterado glandular por cada estado de uma região do Brasil

Os resultados a seguir representam os resultados alterado glandular identificados entre os períodos de 2013 a 2022. Cada gráfico presente na Figura 4.11 representa uma região do Brasil e cada série indica um estado da respectiva região.

A maior concentração de resultados foi encontrada na região Sudeste, com destaque para o estado de São Paulo com a maior quantidade de informações.

As regiões Norte e Centro-Oeste apresentaram os menores resultados, onde, com exceção dos estados do Pará e Goiás, todos apresentaram resultados inferiores a 500 resultados.

Figura 4.10 – Resultados alterado glandular de cada estado em cada uma das regiões do Brasil.

4.11 Resultados, por indicador, em cada região do Brasil

Os resultados dessa seção apresentam os dados quantitativos referentes ao período entre 2013 a 2022. Cada um dos gráficos presentes na Figura 4.11 representam um estado do Brasil e cada uma das séries representam um dos indicadores analisados.

Em todos os casos, os resultados negativos representam montante consideravelmente semelhante em comparação com a quantidade de exames. Tal informação indica que a maioria dos exames realizados indica um resultado negativo.

Pela quantidade de exames representarem a totalidade do conjunto analisado, os resultados alterados acabaram ficando ofuscados. Por isso, a próxima seção apresenta os resultados considerando somente os resultados alterados.

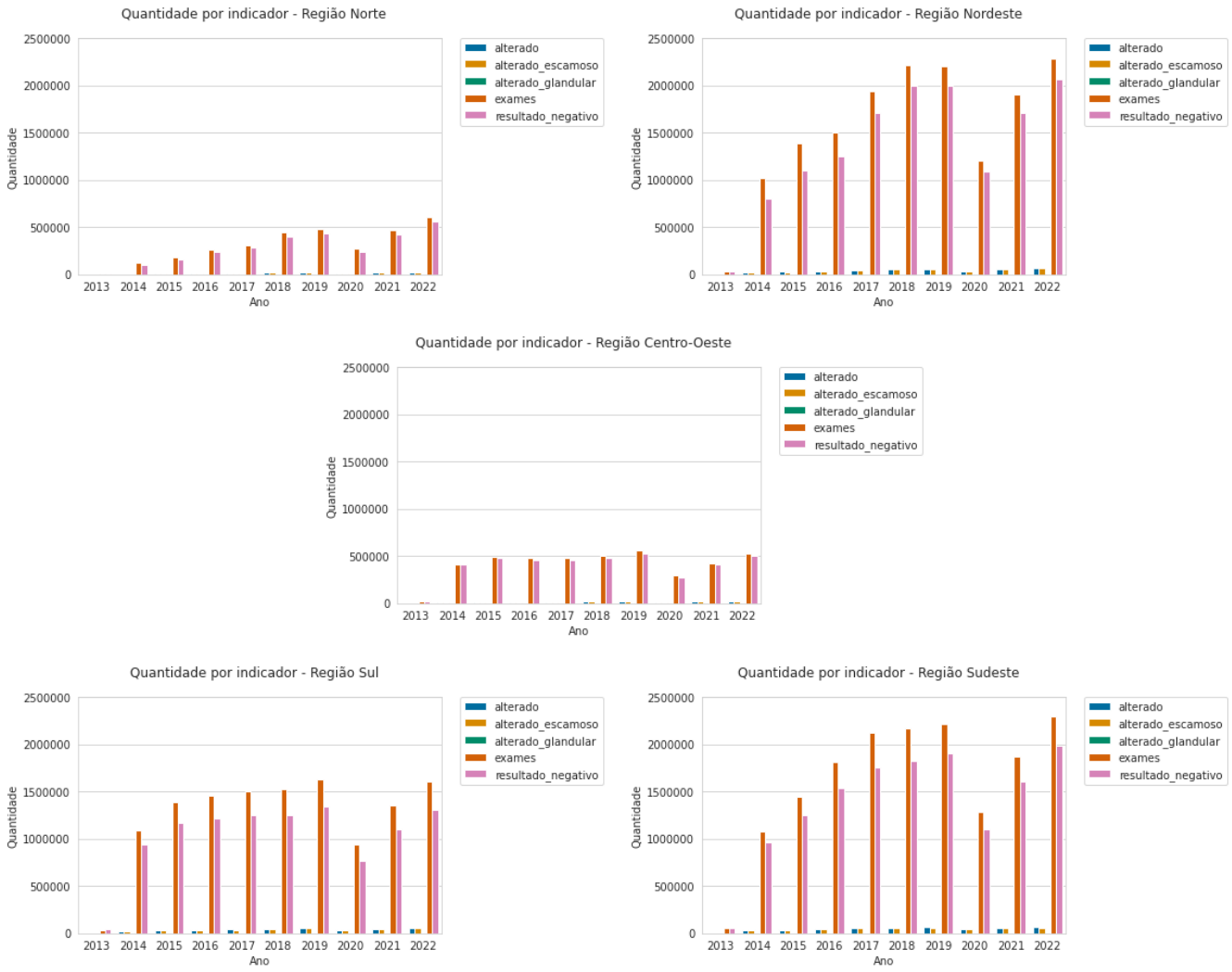


Figura 4.11 – Resultados em cada região do Brasil sobre os indicadores estudados.

4.12 Indicadores de resultados alterados em cada região do Brasil

A última seção desse capítulo apresenta as informações presentes na Figura 4.12, onde cada gráfico representa uma região do Brasil e cada série representa um tipo de indicador. Nessa figura, serão avaliados somente os indicadores alterados, onde inclui-se os resultados alterados, resultados alterado escamoso e resultado alterado glandular.

Pelo indicador de resultados alterados apresentarem a totalidade dos resultados, o mesmo apresentou sempre os maiores montantes entre os indicadores avaliados. Porém, em todos os casos, os resultados alterados escamosos apresentaram dados muito próximos em relação aos resultados alterados, indicando que a maioria dessa totalidade são identificados a partir desse indicador. Sabendo disso, os resultados alterados glandulares representaram graficamente como montantes pouco significantes em relação aos exames indicados como alterados.

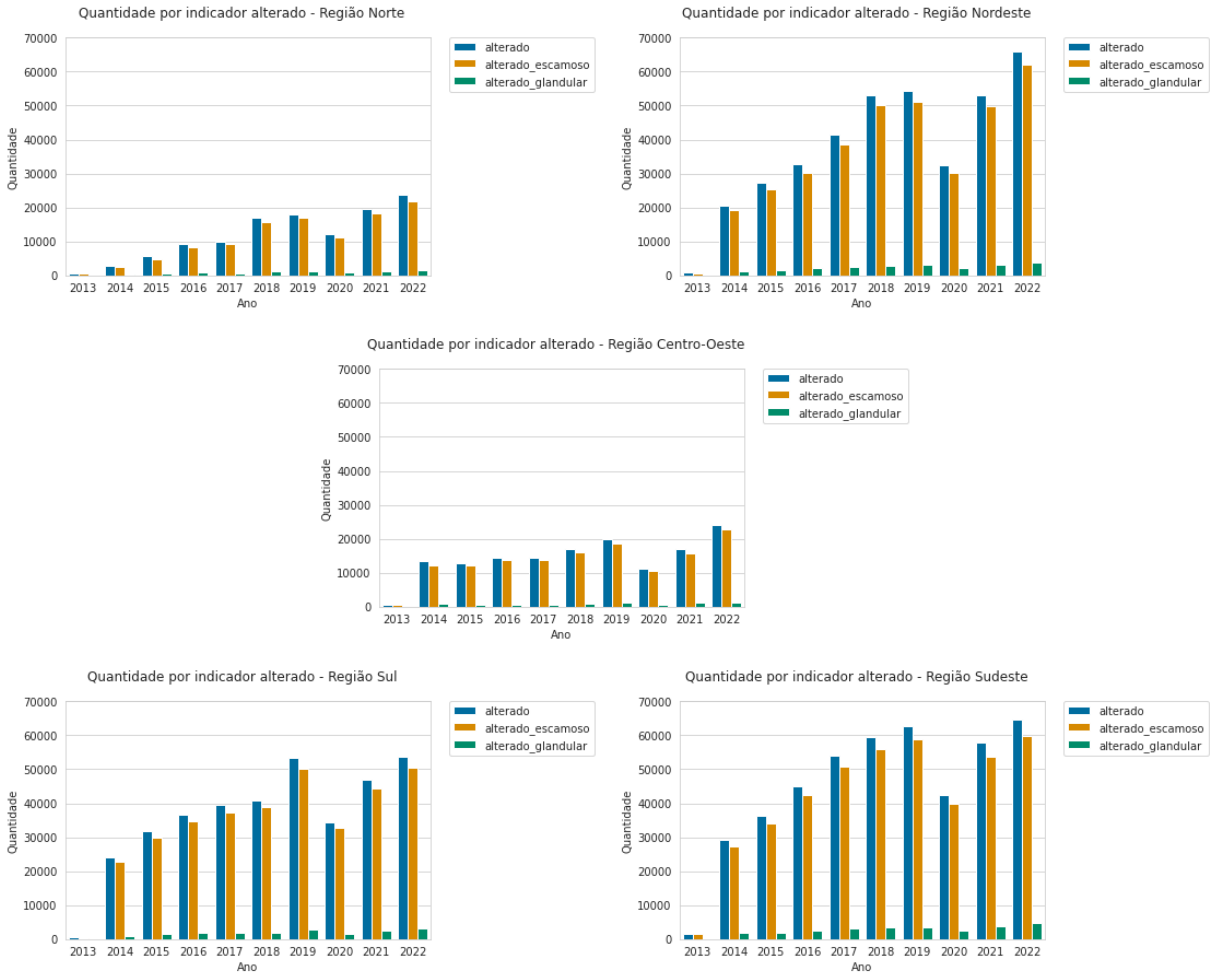


Figura 4.12 – Resultados em cada região do Brasil sobre os indicadores de resultados alterados.

5 Conclusões e Trabalhos Futuros

A utilização de ferramentas de visualização de dados tem como principal objetivo representar visualmente um conjunto de dados para realização de análises e identificação de padrões. Sabendo disso, o presente trabalho analisa dados temporais quantitativos presentes no sistema do SISCAN/DATASUS por meio da elaboração de gráficos e discute as informações sobre a realização de exames de citologia do colo do útero.

O trabalho desenvolvido contempla o processo de obtenção, tratamento e persistência dos dados, a elaboração de consultas de dados que retornam dados já com a utilização de regras de negócio e, por fim, a elaboração de gráficos que representam visualmente as informações quantitativas e temporais obtidas no SISCAN.

Após a descrição da estratégia de visualização de dados elaborada, foram apresentados os gráficos elaborados a partir de dados do SISCAN, de maneira a detalhar como eram os dados anteriores a pandemia da Covid-19, desde o primeiro ano de armazenamento e divulgação aberta dos dados do SISCAN sobre o exame de Papanicolau (2013), até o que deveria acontecer de acordo com projeções e o que de fato aconteceu durante e após o primeiro ano de pandemia, nos anos de 2020 a 2022. Estes dados foram segregados por região do Brasil e estado da unidade federativa.

Notou-se que, independente de região, os dados seguem um padrão: o primeiro ano de divulgação dos dados (2013) apresentam valores quantitativos muito baixos, o que pode ser explicado pelos dados do ano em questão serem representados somente a partir do ano de Junho. Após isso, foi encontrado um padrão onde os resultados tendem aumentar significativamente entre os anos de 2014 a 2016 e, após isso, tenderem a uma taxa de crescimento quase nula, mantendo as quantidades exibidas praticamente iguais e com quedas/aumentos pouco significativos entre os 3 anos anteriores a pandemia (2017 a 2019). No primeiro ano de pandemia, em 2020, destaca-se uma queda brusca em todas as localidades do Brasil, com variações entre 19% a 48%, dependendo da região em comparação ao primeiro ano anterior ao qual a Covid-19 se espalhou por todo o planeta.

Em 2021, notou-se também um padrão de retomada significativa em relação a quantidade de exames, mesmo que ainda estando em uma situação de calamidade pública diante todo o mundo: em relação a 2019, as taxas ficaram entre -23% a 14%. Uma das possibilidades de retomada são o surgimento de pesquisas demonstrando o funcionamento do vírus e maneiras de proteção ao contágio eficientes ou até mesmo indicando a falta de políticas públicas no país para controlar o fluxo da população nas cidades brasileiras.

Na maioria das análises apresentadas, o estado do Rio de Janeiro apresentou o movimento oposto ao aquele que era esperado: a taxa de crescimento dentre os anos da pandemia, de 2020 a

2022, apresenta valores positivos, demonstrando aumento nos resultados ao invés da queda.

A região norte apresentou as menores quantidades de exames de Papanicolau realizados no período, com pontos máximos pouco superiores a meio milhão de exames realizados em um ano. Considerando a taxa de crescimento médio realizado na predição (simulação) de valores, a previsão era chegar em patamares próximos ao primeiro milhão de exames realizados em um único ano (2022). Teve-se o Pará como o estado com mais exames realizados e o Amapá como o estado com menos exames realizados.

A região nordeste apresenta poucas informações ao visualizá-las por estado, porém, por possuir uma grande quantidade de estados em sua região, o seu valor final acaba se tornando inflado e pode persuadir o analista de dados. Teve-se a Bahia como o estado com mais exames realizados e o Piauí como o estado com menos exames realizados.

Já na região centro-oeste foi considerada a segunda região com menos informações relacionadas ao exame de Papanicolau. Apesar disso, os três estados da região apresentaram quantidades de exames semelhantes entre si, com exceção do Distrito Federal, que, além disso, só passou a ser visivelmente relevante nos gráficos elaborados a partir de 2018. Teve-se o Goiás como o estado com mais exames realizados e o Distrito Federal com menos exames realizados.

A região sul manteve-se, desde 2014, com a quantidades entre um milhão e um milhão e meio de exames de Papanicolau realizados em um ano, com picos em 2019 e 2022 tendo, aproximadamente, um milhão e seiscentos mil exames. Apesar disso, levando em consideração os dados simulados, poderia atingir um patamar de quase um milhão e novecentos mil exames em 2022. Teve-se o Paraná como o estado com mais exames realizados e Santa Catarina como o estado com menos exames realizados.

Por fim, a região sudeste apresentou, em suma maioria, o maior resultado dentre todas as regiões do Brasil, indicando a possibilidade de uma maior conscientização da população sobre a importância na realização do exame para identificar e prevenir o câncer de colo do útero. Apesar disso, mesmo apresentando crescimento contínuo nos resultados, a quantidade de exames realizados ainda é pequena comparada com a população total da região, sendo uma das mais populosas do Brasil.

5.1 Trabalhos Futuros

Vislumbram-se as seguintes perspectivas de trabalhos futuros:

- Criação de uma ferramenta gráfica para gestão de gráficos sob demanda do analista de dados;
- Realizar o cruzamento de dados com diversas informações governamentais como, por exemplo, dados populacionais do IBGE;

- Elaborar um fluxo de consumo automatizado dos dados obtidos a partir do SISCAN/DATASUS;
- Criação de uma API para disponibilização pública das informações.

Referências

BRASIL, M. da Saúde do. *Papanicolau (exame preventivo de colo de útero)*. 2011. <https://bvsms.saude.gov.br/papanicolau-exame-preventivo-de-colo-de-utero/>. Acessado em: 19/10/2022.

CHAVES, A. K. M.; RESENDE, I. C. de; SOUZA, M. A. D.; AGULHON, N. G.; GONTIJO, T. B.; ZUQUETTI, V. R. V.; MACHADO, L. C. de S. Impacto da pandemia da COVID-19 no rastreamento do câncer do colo uterino no estado de goiás. *Brazilian Journal of Development*, v. 8, n. 2, p. 12989–12988, 2022.

CIOTTI, M.; CICCOZZI, M.; TERRINONI, A.; JIANG, W.-C.; WANG, C.-B.; BERNARDINI, S. The COVID-19 pandemic. *Critical reviews in clinical laboratory sciences*, Taylor & Francis, v. 57, n. 6, p. 365–388, 2020.

DATASUS. *Nota técnica 3 - SISCAN – EXAMES CITOPATOLOGICOS DO COLO DO ÚTERO por município de residência da paciente*. 2019. <http://tabnet.datasus.gov.br/cgi/SISCAN/doc/nota_tecnica_3_cito_colo_resid.pdf>. Acessado em: 10/03/2023.

GOMIDES, T. G. F. Impacto da pandemia COVID-19 no rastreamento do câncer do colo de útero no município de ouro preto-mg. 2022.

INCA. *Instituto Nacional do Câncer - Câncer do colo do útero*. 2022. <https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/colo-do-utero>. Acessado em: 04/09/2022.

MCKINNEY, W. et al. Pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, Seattle, v. 14, n. 9, p. 1–9, 2011.

NEVES, L. R.; EUSTÁQUIO, V. M.; ARAÚJO, R. L. A influência da COVID-19 no diagnóstico de neoplasias de colo uterino e de mama no brasil. *Facit Business and Technology Journal*, v. 1, n. 34, 2022.

ONCOGUIA, E. *Exames de Rastreamento para Câncer de Colo do Útero*. 2014. <http://www.oncoguia.org.br/conteudo/exames-de-rastreamento-para-cancer-de-colo-do-utero/1284/284/>. Acessado em: 04/03/2023.

OPPEL, A. J.; SHELDON, R. *SQL: a beginner's guide*. [S.l.]: McGraw-Hill, 2008.

SILVA, I. N. de Câncer José Alencar Gomes da. *Diretrizes Brasileiras para o Rastreamento do Câncer do Colo do Útero*. 2016. <https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//diretrizes_para_o_rastreamento_do_cancer_do_colo_do_utero_2016_corrigido.pdf>. Acessado em: 22/03/2023.

TELESSAúDERS. *O que significa metaplasia escamosa imatura no resultado do Papanicolau (CP do colo de útero)?* 2015. <https://ares.unasus.gov.br/acervo/html/ARES/2685/1/SOF%20c%C3%A2ncer%20colo%20Telessa%C3%BAdeRS%20220615.pdf>. Acessado em: 04/03/2023.

WASKOM, M. *An introduction to Seaborn*. 2021. <https://seaborn.pydata.org/tutorial/introduction.html>. Acessado em: 04/09/2022.

WASKOM, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software*, v. 6, n. 60, p. 3021, 2021.

WHO. *Coronavirus disease (COVID-19): How is it transmitted?* 2021. <https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted>. Acessado em: 04/09/2022.

WIRFS-BROCK, A.; EICH, B. Javascript: The first 20 years. *Proc. ACM Program. Lang.*, Association for Computing Machinery, New York, NY, USA, v. 4, n. HOPL, jun 2020. Disponível em: <https://doi.org/10.1145/3386327>.