UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

EVANDRO FERNANDES BARRETO

CIÊNCIA DE DADOS APLICADA À PANDEMIA DO CORONAVÍRUS NO BRASIL, UMA ANÁLISE SOCIOECONÔMICA

BAURU

Julho/2021

EVANDRO FERNANDES BARRETO

CIÊNCIA DE DADOS APLICADA À PANDEMIA DO CORONAVÍRUS NO BRASIL, UMA ANÁLISE SOCIOECONÔMICA

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, campus Bauru.

Orientador: Prof. Dr. João Pedro Albino

BAURU

Julho/2021

Barreto, Evandro Fernandes.

Ciência de dados aplicada à pandemia do Coronavirus no Brasil, uma análise socioeconômica/ Evandro Fernandes Barreto – Julho/2021

37 p.: il. (algumas color.); 30 cm.

Orientador: Prof. Assoc. Dr. João Pedro Albino

Trabalho de Conclusão de Curso–Universidade Estadual Paulista. Faculdade de Ciências, Bacharelado em Ciência da Computação. 2021.

1.Ciência de dados. 2. Análise de Dados. 3. Covid-19. 4. Pandemia. I. Universidade Estadual Paulista. Faculdade de Ciências. II. Ciência de dados aplicada à pandemia do Coronavirus no Brasil, uma análise socioeconômica

Evandro Fernandes Barreto

CIÊNCIA DE DADOS APLICADO À PANDEMIA DO CORONAVÍRUS NO BRASIL, UMA ANÁLISE SOCIOECONÔMICA

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Assoc. Dr. João Pedro Albino
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Profa. Dra. Simone das Graças Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Prof. Dr. Kleber Rocha de Oliveira
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Câmpus Experimental de Rosana
Engenharia de Energia

Bauru, 26 de Julho de 2021.

Agradecimentos

Agradeço aos meus pais, que me apoiaram e me incentivaram nos estudos
Agradeço a todos os meus professores que contribuíram para minha formação ao longo dos anos, em especial do meio acadêmico.
Agradeço aos meus amigos, que passaram e estiveram comigo esses anos.
Agradeço em especial meu orientador.

Resumo

Com a pandemia do novo coronavírus o número de estudos aumentou drasticamente e com eles o volume de dados produzidos e distribuídos. A explosão de dados aliado ao avanço no processamento possibilitou o crescimento da Ciência de Dados como base para auxiliar à tomada de decisões. No Brasil, diversos órgãos publicam e publicaram informações e dados sobre a pandemia de COVID-19, dentre eles o Ministério da Saúde, órgão do poder executivo federal, responsável pela promoção, prevenção e assistência à saúde do país. Os dados publicados por esse órgão foram alvo dos estudos desenvolvidos neste trabalho, assim como os dados socioeconômicos produzidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE), órgão federal responsável por prover informações estatísticas e geográficas do Brasil.

Para este estudo, observou-se que há uma grande quantidade de dados produzidos e ainda em constante produção. Entretanto, apesar das informações serem de fácil acesso, sua análise depende de filtrar várias variáveis e correlacionar com outras. Esse aspecto faz com que apenas quem tem conhecimento de alguma ferramenta de Ciência ou Análise de Dados possa chegar a algum resultado.

Além disso, dentre os dados já divulgados sobre a pandemia, poucos são de fácil interpretação para a maioria da população, gerando distanciamento e descrédito tornando ainda mais difícil o combate à pandemia de COVID-19.

Tendo em vista a problemática citada, esse trabalho propõe analisar e disponibilizar resultados que facilitem a visualização dos dados a respeito da pandemia.

O objetivo principal é o de relacionar os dados do Ministério da Saúde sobre coronavírus no Brasil com os dados socioeconômicos do IBGE e inferir se existe uma relação causal entre a propagação da doença e características da população.

Por fim divulgar o resultado em uma plataforma com mapas e gráficos de fácil interpretação e acesso.

Palavras-chave: Pandemia, COVID-19, Ciência de Dados, Análise de Dados, Dados Geográficos, Dados Socioeconômicos, Visualização de Dados.

Abstract

With the new coronavirus pandemic, the number of studies increased dramatically and with them the volume of data produced and distributed. The explosion of data combined with advances in processing enabled the growth of Data Science as a basis to assist decision making. In Brazil, several agencies publish and publish information and data on the COVID-19 pandemic, among them the Ministry of Health, an agency of the federal executive power, responsible for the promotion, prevention and assistance to health in the country. The data published by this agency were the target of the studies developed in this work, as well as the socioeconomic data produced by the Brazilian Institute of Geography and Statistics (IBGE), the federal agency responsible for providing statistical and geographic information about Brazil. For this study, it was observed that there is a large amount of data produced and still in constant production. However, although the information is easily accessible, its analysis depends on filtering out several variables and correlating them with others. This aspect means that only those who have knowledge of a Science or Data Analysis tool can reach some result. Furthermore, among the data already released on the pandemic, few are easy to interpret for the majority of the population, generating distance and discredit, making the fight against the COVID-19 pandemic even more difficult. In view of the aforementioned problem, this work proposes to analyze and make available results that facilitate the visualization of data regarding the pandemic. The main objective is to relate data from the Ministry of Health on coronavirus in Brazil with socioeconomic data from IBGE and infer whether there is a causal relationship between the spread of the disease and population characteristics. Finally, publish the result on a platform with maps and graphics that are easy to interpret and access.

Key words: Pandemic, COVID-19, Data Science, Data Analysis, Geographic Data, Socioeconomic Data, Data Visualization

Lista de figuras

Lista de quadros

Lista de siglas

DBF	Data Base Field
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IDHM	Índice de Desenvolvimento Humano Municipal
IDHM E	Índice de Desenvolvimento Humano Municipal de Educação
IDHM L	Índice de Desenvolvimento Humano Municipal Longevidade
IDHM R	Índice de Desenvolvimento Humano Municipal de Renda
MS	Ministério da Saúde
SAEDE	Sistema Estadual de Analise de Dados
SUS	Sistema Único de Saúde
SRAG	Síndrome Respiratória Aguda Grave
PIB	Produto Interno Bruto
UF	Unidade Federativa

Sumário

1 Introdução

A produção de dados hoje é algo sem precedentes. Exames médicos são realizados por smartphones, celulares mapeiam a localização das pessoas de hora em hora, agências públicas compartilham grandes bancos de dados com universidades e empresas e assim por diante.

Por isso, entre o monitoramento da quarentena, as curvas epidemiológicas e a busca pela vacina, a história da pandemia que enfrentando é também uma história sobre dados. Será que dá para chamar a pandemia de covid-19 de pandemia da ciência de dados? (Salles,2020).

Com a pandemia do novo coronavírus, o número de estudos aumentou drasticamente e junto o volume de dados produzidos e distribuídos por todos os países com objetivo de ajudar a combater o vírus também sofreu um aumento considerável.

A Ciência de Dados não é recente, porém ganhou maior relevância nos últimos anos com o aumento do processamento dos computadores, permitindo que esses conjuntos de informações sejam analisados, ajudando na tomada de decisão.

Portanto, considerando todos esses aspectos citados anteriormente, esse trabalho tem como objetivo, analisar os dados sobre a pandemia da Covid-19 no Brasil, disponibilizados pelo Ministério da Saúde e relacioná-los com os dados socioeconômico dos municípios brasileiros disponibilizados pelo Instituto Brasileiro de Geografia e Estatística(IBGE) e aplicar técnicas de Ciência de Dados para correlacionar os temas e inferir se existe uma relação e, ao final dos processos disponibilizar graficamente os resultados obtidos.

1.2 Problema

Há uma grande quantidade de dados produzidos e ainda em constante produção sobre a pandemia da COVID-19 e, apesar das informações serem de fácil acesso, sua análise e interpretação dependem de filtrar entre as diferentes variáveis utilizadas e correlacioná-las com outros dados para assim poder deduzir e obter informações.

Tal aspecto faz com que apenas quem tem conhecimento de alguma ferramenta de análise e exploração de dados ou de técnicas de Ciência possa chegar a algum resultado.

Além disso, dos dados já divulgados sobre a pandemia, poucos são de fácil interpretação para a maioria da população, gerando um distanciamento e descrédito tornando ainda mais difícil o a compreensão dos problemas ocasionados pela pandemia e o seu combate.

Tendo em vista a problemática citada, esse trabalho propõe analisar e disponibilizar graficamente os resultados de correlação entre os dados oficiais da pandemia e os dados demográficos dos municípios, visando apoiar a tomada de decisão com relação à pandemia.

1.3 Justificativa

Apesar de ser um conhecimento prévio e histórico, a população em geral se vê sem explicações e dados que justifiquem as ações de combate ao vírus tomadas nos diversos Estados e municípios brasileiros, proporcionando um ambiente de dúvida e incerteza e com isso prejudicando a eficiência de tais ações.

1.4 Objetivos

O objetivo do trabalho é relacionar os dados do coronavírus no Brasil, produzidos pelo Ministério da Saúde, com os dados socioeconômicos da população brasileira, produzidos pelo IBGE, inferir se existe uma relação causal e por fim divulgar o resultado em uma plataforma com mapa e gráficos de fácil interpretação e acesso.

2 Fundamentação Teórica

Nos subtópicos deste capítulo, são apresentados alguns conceitos utilizados neste trabalho.

2.1 Ciência de Dados

Ciência de Dados é um estudo muito disciplinado com relação aos dados e demais informações inerentes à empresa e as visões que cercam um determinado assunto. Em resumo é uma ciência que visa estudar os dados; seu processo de captura, transformação, geração e, posteriormente, análise.

A Ciência de Dados envolve diversas disciplinas, afirma Coelho (2020) e afeta de forma acadêmica ou não, pesquisas aplicadas em diversos domínios, como a tradução automática, o reconhecimento de voz e motores de dispositivos de busca.

Mas não só isto, ela impacta também a economia digital, a informática médica nos cuidados com a saúde e ainda influencia fortemente a economia, os negócios e as finanças. (SILVEIRA, 2016). Do ponto de vista empresarial, a Ciência de Dados (ou *Data Science*) tornou-se uma parte vital da inteligência competitiva, um campo emergente que engloba uma série de atividades, como mineração e análise de dados.

2.2 Análise de Dados

Um dos fundadores da análise de dados, Tukey (1961) define seus procedimentos como:

> *Procedimentos para a análise de dados, técnicas de interpretação dos resultados de tais procedimentos, formas de planejar a coleta de dados para tornar sua análise mais fácil, mais precisa ou mais exata, e todos os mecanismos e resultados das estatísticas (matemáticas) que se aplicam à análise de dados.* (Tukey, 1961).

A análise de dados é a arte de transformar dados em conhecimentos e *insights - acontecimento cognitivo que pode ser associado a vários fenômenos podendo ser sinônimo de compreensão, conhecimento, intuição* - relevantes. Ou seja, comparar e agregar as informações brutas para entender o que os dados nos dizem. (ESCOLADEDADOS, 2019).

Dentro dessa área existem subdivisões que a separam em quatro partes:

a. **Análise Descritiva:** Consiste na descrição das principais características de um conjunto de dados, listando e resumindo valores. Nessa subdivisão busca-se uma percepção mais dinâmica, sendo possível aplicar operações como a média, mediana, moda, mínima, máxima, porcentagem e frequência.
b. **Análise Preditiva:** Nessa etapa, utiliza-se do acúmulo histórico de informações e dados para realizar previsões sobre eventos futuros. O objetivo é fazer projeções mais sólidas

do futuro, a partir disso decisões mais adequadas são tomadas de acordo com as expectativas.

c. **Análise Prescritiva:** Análise voltada para projeções, usando resultados de outras análises como objeto de tomada de decisões mais específicas, como por exemplo liberação de crédito automático.
d. **Análise Exploratória:** Voltada para análise de conjuntos, usando métodos visuais como gráficos e ferramentas de visualização de dados, visando maximizar a obtenção de informações nessas estruturas, detectar comportamentos anômalos, testar hipóteses ou determinar o número ótimo de variáveis.

3 Metodologia

Neste capítulo serão descritos os métodos e técnicas de desenvolvimento utilizados.

3.1 Base de Dados

3.1.1 Base de Dados do Ministério da Saúde

O Ministério da Saúde é um órgão do Poder Executivo Federal responsável pela administração e manutenção da saúde pública no Brasil.

Criado em 1930 durante o governo de Getúlio Vargas com o nome de" Ministério dos Negócios da Educação e Saúde Pública", foi apenas em 1953 que houve o desmembramento para Ministério da Saúde.

Com o fim da ditadura na década de 1980 e a consequente Constituição Federal, que determinou ser dever do Estado garantir a saúde de toda a população, criando o Sistema Único de Saúde (SUS), que realiza desde atendimentos primários até procedimentos mais complexos.

Com SUS toda população Brasileira passou a ter Direito à Saúde gratuita, junto a isso, acesso a serviços de vigilância sanitária, vigilância epidemiológica, vigilância ambiental, além de fundações e institutos de pesquisa acadêmica e científica, que nos dias atuais esses serviços disponibilizam dados que formam a plataforma de Dados Abertos chamada de OpenDataSUS.

OpenDataSUS é uma plataforma de dados abertos do Ministério da Saúde que busca disponibilizar informações sanitárias que possam ajudar na tomada de decisão e elaboração de programas de saúde pública, disponibilizando dados sobre a pandemia como registro de ocupação hospitalar, distribuição de equipamentos e o objeto deste trabalho os Bancos de Dados de Síndrome Respiratória Aguda Grave (SRAG) dos anos de 2020 e 2021 em formato tabulado (arquivo .csv).

3.1.2 Base de Dados do Instituto Brasileiro de Geografia e Estatística (IBGE)

O Instituto Brasileiro de Geografia e Estatística (IBGE) é um instituto da administração federal, criado em 1936 tendo como função prover dados e informações do País, dentre eles análises estatísticas, geográficas, ambientais e censos, que atendem e cobre os mais diversos segmentos da sociedade civil bem como órgãos governamentais, federais , estaduais e municipais.

Como objeto deste trabalho foram usados os dados que descrevem as coordenadas em latitude e longitude de cada município, disponibilizadas pelo IBGE e o Índice de Desenvolvimento Humano Municipal (IDHM).

O Índice de Desenvolvimento Humano (IDH) é uma medida resumida do progresso a longo prazo usando três indicadores do desenvolvimento humano: renda, educação e saúde. O objetivo da criação do IDH foi o de oferecer um contraponto a outro indicador muito utilizado, o Produto Interno Bruto (PIB) per capita, que considera apenas a dimensão econômica.

O Índice de Desenvolvimento Humano Municipal é um subconjunto do IDH que avalia o desenvolvimento de cada município brasileiro calculado com dados indiretos do Censo Demográfico do IBGE baseado em 180 indicadores. O índice varia de 0 a 1, quanto mais próximo do número 1 maior é o desenvolvimento, para este trabalho foram usados os indicadores de renda (IDHM R), longevidade (IDHM L), educação(DHM E) e o conjunto dos três formando o IDHM obtidos em 2010.

3.2 Ferramentas

3.2.1 Google Colabolatory

Google Colabolatory é um ambiente de serviço nuvem gratuito hospedado pelo Google que permite escrever códigos no navegador sem necessidade de configuração.

A Ferramenta permite executar códigos e compartilhá-los de maneira fácil, além de incluir imagens e textos.

Todo o poder computacional utilizado para executar o software que você escreve é fornecido pela nuvem de computadores da Google. Dando gratuitamente ao usuário a possibilidade de processar uma quantidade bem grande de dados(Roveda,2019).

3.2.2 Python

Linguagem de programação de alto nível orientada a objeto, tipado dinamicamente, publicada em 1991 por Guido Van Rossum e atualmente possui um sistema de desenvolvimento aberto.

Devido a sintaxe simples, um código escrito em Python geralmente é menor e mais legível se comparado com outras linguagens, como C++ ou PHP. Há menos exigências “gramaticais”, como parênteses em estruturas de seleção ou ponto-e-vírgula no fim da linha, e o código é estruturado com base em espaços em brancos. (Melo,2021)

Python foi a linguagem escolhida para esse trabalho, uma vez que se torna cada mais popular no meio acadêmico e no mercado, além de possuir diversas bibliotecas que facilitam e flexibilizam o uso,[...] você pode encontrar uma grande variedade bibliotecas de Ciência de dados (como por exemplo: NumPy, SciPy, StatsModels, scikit-learn, pandas, etc.), que estão em crescimento exponencial.

Restrições (em métodos de otimização / funções) que estavam faltando um ano atrás já não são um problema e você pode encontrar uma solução robusta adequada, que funciona de forma confiável. (Matos,2019).

3.2.2.1 Pandas

Biblioteca para a linguagem Python, para a manipulação e análise de dados, oferecidos estruturas e operações para manipular tabelas numéricas e séries temporais, publicada em 2008, tornando de código aberto em 2009.

3.2.2.2 DataFrame

Estrutura bidimensional com tamanho mutável, contendo eixos (linha e colunas) rotulados sendo a principal estrutura de dados do pandas, a Figura 1 abaixo ilustra as informações contendo informações sobre municípios.

Figura 1 - Exemplo de DataFrame

	CD_MUN	NM_MUN	SIGLA_UF
0	1100015	Alta Floresta D'Oeste	RO
1	1100023	Ariquemes	RO
2	1100031	Cabixi	RO
3	1100049	Cacoal	RO
4	1100056	Cerejeiras	RO

Fonte: Elaborado pelo autor.

3.2.2.3 GeoDataFrame

GeoDataFrame é uma estrutura similar ao DataFrame, que contém duas outras estruturas,GeoSeries e DataFrame.

GeoSeries é essencialmente um vetor em que cada entrada corresponde a um conjunto e forma, esse conjunto pode ser por exemplo um que continue um País ou Estado, transformando dados de longitude e latitude em dados geométricos. Na figura 2 contém um exemplo de GeoDataframe com polígonos que formam os municípios brasileiros.

Figura 2 - Exemplo de GeoDataFrame

NM_MUN	SIGLA_UF	AREA_KM2	geometry
Alta Floresta D'Oeste	RO	7067.025	POLYGON ((-62.22630 -11.89037, -62.20670 -11.8...
Ariquemes	RO	4426.571	POLYGON ((-63.58751 -9.84984, -63.58715 -9.849...
Cabixi	RO	1314.352	POLYGON ((-60.71834 -13.39058, -60.70904 -13.3...
Cacoal	RO	3792.892	POLYGON ((-61.50114 -11.30119, -61.50104 -11.2...
Cerejeiras	RO	2783.300	POLYGON ((-61.51346 -13.28575, -61.51534 -13.2...

Fonte: Elaborado pelo autor.

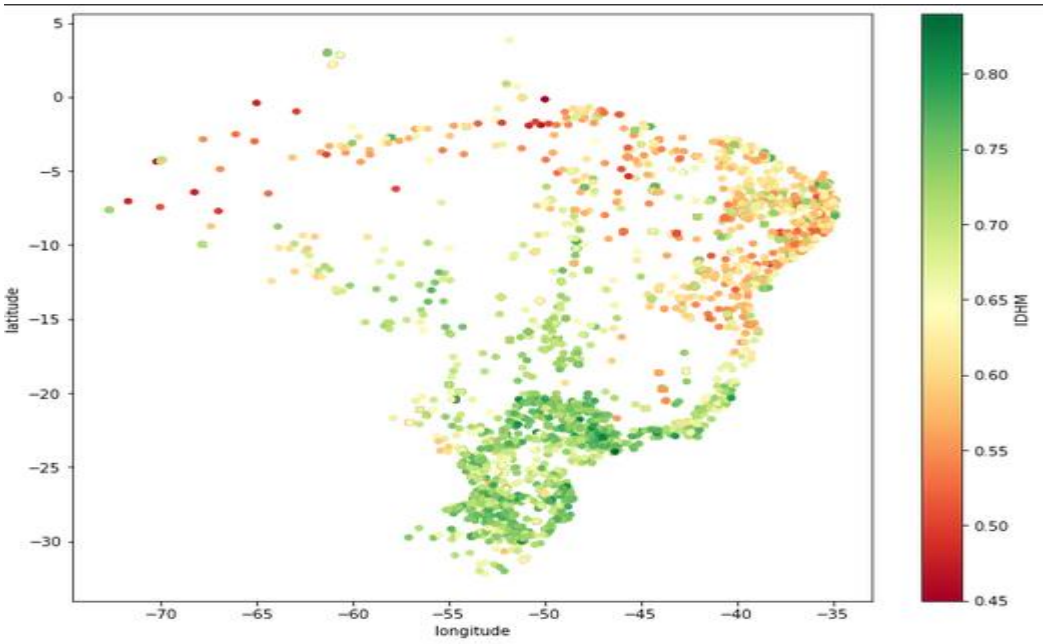
3.2.2.4 Matplotlib

A Biblioteca para criação de gráficos e visualização de dados em geral da linguagem Python junto com a sua extensão numérica, Numpy, foi criada pelo biólogo e neurocientista americano John D. Hunter para ser uma alternativa aberta ao Matlab, um software de alta performance voltado para cálculos numéricos, matrizes, processamentos de sinais e construção de gráficos.

Numpy pacote que suporta arrays –*estruturas usadas para armazenar diferentes valores em uma única variável* - e matrizes multidimensionais, possuindo uma larga coleção de funções matemáticas para trabalhar com estas estruturas.

A Figura 3 exemplifica um gráfico geográfico usando as informações do IDHM , que foram usadas no projeto.

Figura 3 - Exemplo de Gráfico Geográfico utilizando Matplotlib



Fonte: Elaborado pelo autor.

4 Desenvolvimento

O processo de desenvolvimento foi adotado um processo dividido em quatro etapas:

4.1 Aquisição de dados

Nessa etapa foi realizada a pesquisa de bases de dados e fontes para usar no trabalho.

Como resultado desse processo foi a obtenção de cinco bases de dados diferentes que foram utilizadas.

1. Base de Dados da Notificação de SRAG 2020 - base de dados contém todos os registros do SUS sobre incidência de Síndrome Aguda Grave do ano de 2020.
2. Base de Dados da Notificação de SRAG 2020 - base de dados contém todos os registros do SUS sobre incidência de Síndrome Aguda Grave do ano de 2021.
3. Base de Dados IDHM conjunto de dados com indicadores usados para compilar os índices de IDHM, IDHM longevidade, IDHM Educação e IDHM de Renda.
4. Base contendo as informações de polígonos de cada Município Brasileiro.
5. Base contendo as informações de Latitude de Longitude de cada Município.

Quadro 1 - Exemplo de Dicionário da Base de SRAG

			notificação.		
2-Data de 1ºs sintomas	Date DD/MM/AAAA		Data de 1º sintomas do caso.	**Campo Obrigatório** Data deve ser <= a data da digitação e data do preenchimento da ficha de notificação	**DT_SIN_PRI**
Semana Epidemiológica dos Primeiros Sintomas	Varchar2(6)		Semana Epidemiológica do início dos sintomas.	**Campo Interno** Calculado a partir da data dos Primeiros Sintomas. (SS)	**SEM_PRI**
3-UF	Varchar2(2)	Tabela com código e siglas das UF padronizados pelo IBGE.	Unidade Federativa onde está localizada a Unidade Sentinela que realizou a notificação.	**Campo Obrigatório** Se usuário que está digitando a ficha for de nível: ▪ Unidade Sentinela - o campo é preenchido automaticamente pelo sistema com a UF, município e unidade onde está cadastrado o usuário. ▪ Municipal – o campo é preenchido automaticamente pelo sistema com a UF e município onde está cadastrado o usuário. ▪ Estadual – o campo é preenchido automaticamente pelo sistema com a UF do usuário. ▪ Federal - abre tabela com todas as UF que possuam unidades sentinelas cadastradas no sistema.	**SG_UF_NOT**

Fonte: Elaborado pelo autor.

Durante o processo optou-se por utilizar a base de Síndrome Respiratória,pois com o passar do ano de 2020 para 2021, o número de registros publicados se manteve mais próximos dos dados publicados pela mídia e pelo Ministério da Saúde e outros órgãos, em relação a Base de Dados que registra as mortes provocadas pelo vírus.

4.1.1 Função read_csv e read_file

Função da biblioteca Pandas permite fazer a leitura das base de dados em formato .csv ou file se o arquivo estiver no disco, os paramentos sep indica o caractere que separa as colunas, o encoding indica a codificação de dos caracteres e decimal indica como será a notação de números decimais ao ser lido, por padrão a notação é separação por pelo caractere 'ponto'.

As figuras Figura 4, Figura 5, Figura 6 e Figura 7 mostram o código usado para ler os arquivos de dados pela biblioteca.

Figura 4 - Base de dados SRAG 2020

```
uri = "https://s3-sa-east-1.amazonaws.com/ckan.saude.gov.br/SRAG/2020/INFLUD-21-06-2021.csv"
srag2020 = pd.read_csv(uri,sep=";")
```

Fonte: Elaborado pelo autor.

Figura 5 - Base de dados SRAG 2021

```
uri = "https://opendatasus.saude.gov.br/dataset/9f76e80f-a2f1-4662-9e37-71084eae23e3/resource/42bd5e0e-d61a-4359-942e-ebc83391a137/download/influd21-21-06-2021.csv"
srag2021 = pd.read_csv(uri,sep=";")
```

Fonte: Elaborado pelo autor.

Figura 6 - Base de dados IDHM

```
uri = "/content/drive/MyDrive/BR_Municipios/IDM-BRasil.csv"
idhm = pd.read_csv(uri,sep=";",encoding='ISO-8859-1', decimal=',')
```

Fonte: Elaborado pelo autor.

Figura 7 - Base de dados Polígonos Municipais

```
geo_df = gpd.read_file(arquivo_BR_municipios_shapefile)
geo_df
```

Fonte: Elaborado pelo autor.

Figura 8 - Base de dados Latitude e Longitude dos Municípios

```
uri = "https://raw.githubusercontent.com/kelvins/Municipios-Brasileiros/main/csv/municipios.csv"
municipios = pd.read_csv(uri,sep=",")
municipios
```

Fonte: Elaborado pelo autor.

4.2 Análise dos Dados

Nessa etapa onde foi realizada a análise das variáveis existentes nas bases de dados, algumas das bases possuem documentações chamadas de "dicionário", onde há uma explicação de cada campo e coluna existente.

O resultado desse processo foi a constatação de quais colunas seriam usadas de acordo com análise exploratória e quais informações nessas variáveis serão usadas, explícitas no Quadro 2 e Quadro 3.

Quadro 2 - Campos utilizados da Base SRAG

Nome do Campo	Descrição	Características	DBF
1-Data do preenchimento da ficha de notificação	Data de preenchimento da ficha de notificação	Campo Obrigatório Data deve ser <= a data da digitação	DT_NOTIFIC
4- Município IBGE	Município onde está localizada a Unidade Sentinela que realizou a notificação.	Campo Obrigatório	ID_MUNICIP
1 - Sexo	1- Masculino 2- Feminino 9- Ignorado	Campo Obrigatório	CS_SEXO
15- Raça/Cor	1- Branca 2- Preta 3- Amarela 4- Parda 5- Indigena 9- Ignorado	Campo Obrigatório	CS_RACA
23- UF	Tabela com código e siglas das UF padronizadas pelo IBGE	Campo Obrigatório Se o campo 25-País Brasil	SG_UF
75 - Classificação Final do Caso	1 - SRAG por influenza 2 - SRAG por outro vírus 3 - SRAG por outro agente etiológico 4 - SRAG não especificado 5 - SRAG por covid	Campo Obrigatório	CLASSI_FIN
77 - Evolução do Caso	1 - Cura 2 - Óbito 3 - Óbito por outras causas 9 - Ignorado	Campo Essencial	EVOLUCAO

Fonte: Elaborado pelo autor.

Quadro 3 - Colunas que serão utilizadas em na base de IDHM

Município	IDHM	IDHM_E	IDHM_L	IDHM_R
ALTA FLORESTA D'OESTE	0.641	0.526	0.763	0.657
ARIQUEMES	0.702	0.600	0.806	0.716
CABIXI	0.650	0.559	0.757	0.650
CACOAL	0.718	0.620	0.821	0.727
CEREJEIRAS	0.692	0.602	0.799	0.688

Fonte: Elaborado pelo autor.

4.2.1 Função shape()

Função da biblioteca Pandas permite visualizar a quantidade de linhas e colunas, assim podemos ver a quantidade de dados brutos em cada base de dados.

Na Figura 9 e Figura 10 são mostrados o resultado da função shape(), aplicadas na tabela de SRAG de 2020 e 2021.

Figura 9 - Exemplo da função shape aplicada a SRAG 2020

```
srag2020.shape
```

```
(1196025, 154)
```

Fonte: Elaborado pelo autor.

Figura 10 - Exemplo código da função shape aplicada a SRAG 2021

```
srag2021.shape
```

```
(1185228, 162)
```

Fonte: Elaborado pelo autor.

4.3 Filtragem dos Dados

Nessa etapa são usadas as informações adquiridas na segunda etapa para retirar das bases todas as variáveis que não serão necessárias para o desenvolvimento do projeto.

Tendo em mãos, os resultados da Etapa 2, foram aplicados os filtros em cada uma das bases, tendo como resultado bases com variáveis resultantes.

4.3.1 Função Filter()

Função filter do pandas recebe como parâmetro as colunas deseja manter no data frame ou passar para nome dataframe, dessa forma usamos a função para manter apenas as colunas desejadas, na figura 11 e 12 é demonstrado o código sendo aplicado.

Figura 11 - Função Filter() aplicados nas bases de SRAG 2020 e 2021

```
filter = ['EVOLUCAO','CLASSI_FIN','SG_UF','ID_MUNICIP','CS_RACA','CS_SEXO','DT_NOTIFIC',]
srag2020 = srag2020.filter(items=filter)
srag2021 = srag2021.filter(items=filter)
```

Fonte: Elaborado pelo autor.

Figura 12 - Função Filter() aplicada na base de IDHM

```
filter = ['Município','IDHM','IDHM_E','IDHM_L','IDHM_R']
idhm = idhm.filter(items=filter)
idhm
```

Fonte: Elaborado pelo autor.

Após as filtragens das colunas foi aplicado uma segunda filtragem agora por resultados desejados para o trabalho nas bases de dados de Síndrome Respiratória, seguindo a documentação dos dicionários de dados, sendo selecionadas apenas as ocorrências comprovadas de morte e teste positivo para coronavírus, na figura 13 demonstra como o código foi aplicado.

Figura 13 - Filtragem por resultados de mortes com teste Positivo para Covid-19.

```
sgrag2020filtro = srag2020[(srag2020['EVOLUCAO'] == 2) & (srag2020['CLASSI_FIN'] == 5)]
sgrag2021filtro = srag2021[(srag2021['EVOLUCAO'] == 2) & (srag2021['CLASSI_FIN'] == 5)]
```

Fonte: Elaborado pelo autor.

4.4 Análise das Informações e Busca por Correlações

Nessa etapa, busca-se relacionar os dados resultantes das etapas anteriores, sendo esses ilustrados na Figura 14, Figura 15, Figura 16 e Figura 17.

Figura 14 - Base de dados SRAG 2020 resultante

	EVOLUCAO	CLASSI_FIN	SG_UF	ID_MUNICIP	CS_RACA	CS_SEXO	DT_NOTIFIC
899494	2.0	5.0	CE	FORTALEZA	4	F	01/02/2021
364665	2.0	5.0	AM	MANAUS	4	M	01/02/2021
365724	2.0	5.0	CE	QUIXERAMOBIM	9	M	01/02/2021
17069	2.0	5.0	BA	SALVADOR	4	M	01/02/2021
367392	2.0	5.0	PA	CASTANHAL	4	M	01/02/2021

Fonte: Elaborado pelo autor.

Figura 15 - Base de dados SRAG 2021 resultante

	EVOLUCAO	CLASSI_FIN	SG_UF	ID_MUNICIP	CS_RACA	CS_SEXO	DT_NOTIFIC
887670	2.0	5.0	ES	VITORIA	4.0	M	01/01/2021
1122744	2.0	5.0	SP	SALTO	9.0	F	01/01/2021
793052	2.0	5.0	PR	CURITIBA	4.0	M	01/01/2021
769343	2.0	5.0	SP	SAO BERNARDO DO CAMPO	1.0	F	01/01/2021
1122814	2.0	5.0	SE	ARACAJU	4.0	F	01/01/2021

Fonte: Elaborado pelo autor.

Figura 16 - Base de dados geométricos dos municípios

	NM_MUN	SIGLA_UF	geometry	CS_RACA	CS_SEXO	DT_NOTIFIC	ID_MUNICIP
0	ABADIA DOS DOURADOS	MG	POLYGON ((-47.61843 -18.30777, -47.62127 -18.3...	1.0	F	11/07/2020	5
1	ABAETETUBA	PA	POLYGON ((-49.13483 -1.68729, -49.13130 -1.684...	4.0	F	05/06/2020	81
2	ABAIARA	CE	POLYGON ((-39.10138 -7.34337, -39.10175 -7.341...	1.0	F	01/07/2020	4
3	ABEL FIGUEIREDO	PA	POLYGON ((-48.52422 -5.01698, -48.52462 -5.017...	4.0	M	10/05/2021	4
4	ABELARDO LUZ	SC	POLYGON ((-52.41878 -26.58662, -52.42002 -26.5...	4.0	M	27/02/2021	7

Fonte: Elaborado pelo autor

Figura 17 - Base de dados latitudinais dos municípios

	codigo_ibge	nome	latitude	longitude	capital	codigo_uf
0	5200050	Abadia de Goiás	-16.75730	-49.4412	0	52
1	3100104	Abadia dos Dourados	-18.48310	-47.3916	0	31
2	5200100	Abadiânia	-16.19700	-48.7057	0	52
3	3100203	Abaeté	-19.15510	-45.4444	0	31
4	1500107	Abaetetuba	-1.72183	-48.8788	0	15

Fonte: Elaborado pelo autor.

Como resultado da análise de todos as tabelas exposta, foi realizada a concatenação e junção dessas bases usando as funções de concat() para concatenar as tabelas de Síndrome Respiratória e merge() nas demais, relacionando elas com a coluna correspondendo ao nome de município de cada uma.

A Figura 18 ilustra o código da função *concat()* da biblioteca do pandas e a Figura 19 ilustra a junção da tabela geométrica com a resultante da Figura 18.

Figura 18 - Concatenando as tabelas de SRAGs.

```
srag = pd.concat([sgrag2020filtro,sgrag2021filtro])
```

	EVOLUCAO	CLASSI_FIN	SG_UF	ID_MUNICIP	CS_RACA	CS_SEXO	DT_NOTIFIC
887670	2.0	5.0	ES	VITORIA	4.0	M	01/01/2021
1122744	2.0	5.0	SP	SALTO	9.0	F	01/01/2021
793052	2.0	5.0	PR	CURITIBA	4.0	M	01/01/2021
769343	2.0	5.0	SP	SAO BERNARDO DO CAMPO	1.0	F	01/01/2021
1122814	2.0	5.0	SE	ARACAJU	4.0	F	01/01/2021
...	...	...	...	...	...	...	...
467392	2.0	5.0	PA	MARABA	2.0	F	31/12/2020
784161	2.0	5.0	MT	CUIABA	4.0	F	31/12/2020
931119	2.0	5.0	SC	FLORIANOPOLIS	1.0	F	31/12/2020
311388	2.0	5.0	SP	BRAGANCA PAULISTA	1.0	M	31/12/2020
219954	2.0	5.0	SP	BAURU	1.0	M	31/12/2020

472634 rows × 7 columns

Fonte: Elaborado pelo autor.

O resultado dessa concatenação é uma tabela com registros apenas de mortes por covid, totalizando 472.634 registros em 7 colunas.

Figura 19 - Código usando a função merge() correlacionando duas bases.

```
geo_df['NM_MUN'] = geo_df['NM_MUN'].str.upper()
sragNacional = geo_df.merge(srag,left_on='NM_MUN',right_on='ID_MUNICIP')
```

Fonte: Elaborado pelo autor.

Com três das cincos tabelas tratadas, nessa parte da etapa de correlação foi agrupado os registros por municípios mantidos nas outras colunas e guardando os valores absolutos na coluna de ID_MUNICIP, como resultado temos uma tabela com o total de ocorrência por cada município indicado pela string ‘count’,como mostra a Figura 20.

Figura 20 - Exemplificando o código GroupBy e Agg().

```
dfSrag = sragNacional.groupby(['ID_MUNICIP'])\
.agg({'NM_MUN':'first','SIGLA_UF':'first','geometry':'first','CS_RACA':'first','CS_RACA':'first','CS_SEXO':'first','DT_NOTIFIC':'first','ID_MUNICIP':'count',})\
.reset_index(drop=True)
dfSrag
```

Fonte: Elaborado pelo autor.

O efeito desse agrupamento é uma tabela com dados geográficos de cada município e números absolutos de registo de covid-19, podendo ser vista na Figura 21.

Figura 21 -Tabela resultante do agrupamento.

	NM_MUN	SIGLA_UF	geometry	CS_RACA	CS_SEXO	DT_NOTIFIC	ID_MUNICIP
0	ABADIA DOS DOURADOS	MG	POLYGON ((-47.61843 -18.30777, -47.62127 -18.3...	1.0	F	11/07/2020	5
1	ABAETETUBA	PA	POLYGON ((-49.13483 -1.68729, -49.13130 -1.684...	4.0	F	05/06/2020	81
2	ABAIARA	CE	POLYGON ((-39.10138 -7.34337, -39.10175 -7.341...	1.0	F	01/07/2020	4
3	ABEL FIGUEIREDO	PA	POLYGON ((-48.52422 -5.01698, -48.52462 -5.017...	4.0	M	10/05/2021	4
4	ABELARDO LUZ	SC	POLYGON ((-52.41878 -26.58662, -52.42002 -26.5...	4.0	M	27/02/2021	7

Fonte: Elaborado pelo autor.

Por fim, unimos usando merge() a tabela resultante com os índices de Desenvolvimento humano, ainda correlacionando as colunas por município, que serão usados para comparar o número de registro, o código é mostrado na Figura 22.

Figura 22 - Código merge() com os indicadores de IDHM

```
df = dfSrag.merge(idhm,left_on='NM_MUN',right_on='Município')
```

Fonte: Elaborado pelo autor.

Ao final dessa etapa um dataframe - *uma estrutura de dados rotulada bidimensional com colunas de tipos potencialmente diferentes* - resultante da concatenação e junção de cincos outros que sofreram transformações necessárias para esse trabalho.

4.4.1 Função Plot(), Subplot() e Scatter()

As funções do matplotlib *plot(), subplot() e scatter(),* além de outras incluídas na biblioteca , possibilitam a geração de gráficos temáticos, dinâmicos e flexíveis e customizados, sendo possível alterar cores, inserir títulos e marcações.

A Figura 23 representa o código responsável por gerar um mapa geográfico usando a biblioteca Matplotlib para gerar um gráfico de dispersão, usando a função plot.scatter() das mortes por coronavírus por município, usando a latitude e longitude dos mesmo de referência e a função subplot(), permite criar sub gráficos,possibilitando a visualizar relações entre diferentes gráficos.

Figura 23 - Código usando Plot(), Subplot() e Scatter()

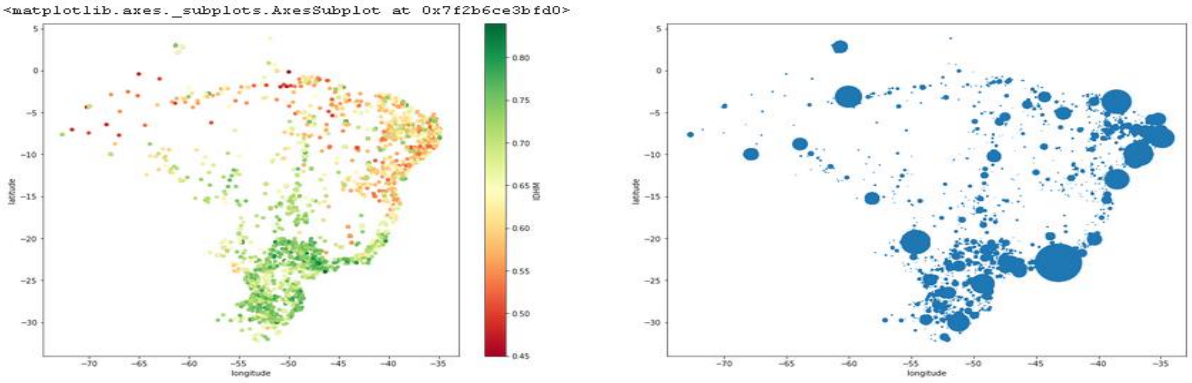
```
fig, ax = plt.subplots(2,figsize=(38,20))
df.plot.scatter(y='latitude',x='longitude',cmap='RdYlGn',c='IDHM',ax=ax[0])
df.plot.scatter(y='latitude',x='longitude',cmap='RdYlGn',marker='o',s=df['ID_MUNICIP']/10,ax=ax[1])
```

Fonte: Elaborado pelo autor.

No código da Figura 24 acima, são criados dois mapas de tamanho 38x20, onde a posição de cada mapa é ditado por um vetor de duas dimensões ax[1], cada posição recebe uma coluna de referência, respectivamente, ax[0] com IDHM e ax[1] de ID_MUNICIP, o parametro “s” define o tamanho da

marcação, como pode ser visto o segundo mapa recebe um tamanho dinâmico da marcação variando de acordo com a quantidade de registros dividido por dez, o resultado é mostrado abaixo na Figura 24.

Figura 24 - IDHM e Mortes por Covid-19



Fonte: Elaborado pelo autor.

5 Resultado

Como resultado deste trabalho, foi obtida a transformação dos dados de cinco tabelas logicamente agregadas em um único *dataframe*, onde estão presentes todas as informações necessárias para gerar visualização gráfica e relacionar os dados socioeconômicos com os dados da pandemia. Por conta disso foi possível criar uma função genérica na qual, variando apenas o parâmetro “Estado”, é possibilitada a visualização dos indicadores pandêmicos graficamente pela UF escolhida

O código gerado para função genérica citada anteriormente está representado na Figura 25, nela uma matriz [3,3], contendo cinco gráficos diferentes com indicadores de IDHM e IDHM por renda, longevidade e educação, para facilitar a visualização, os dados de morte da pandemia, representados pela sigla SIM (Sistema de informação sobre Mortalidade), foram mantidos entre os indicadores.

Figura 25 - Código para indicadores por Estado

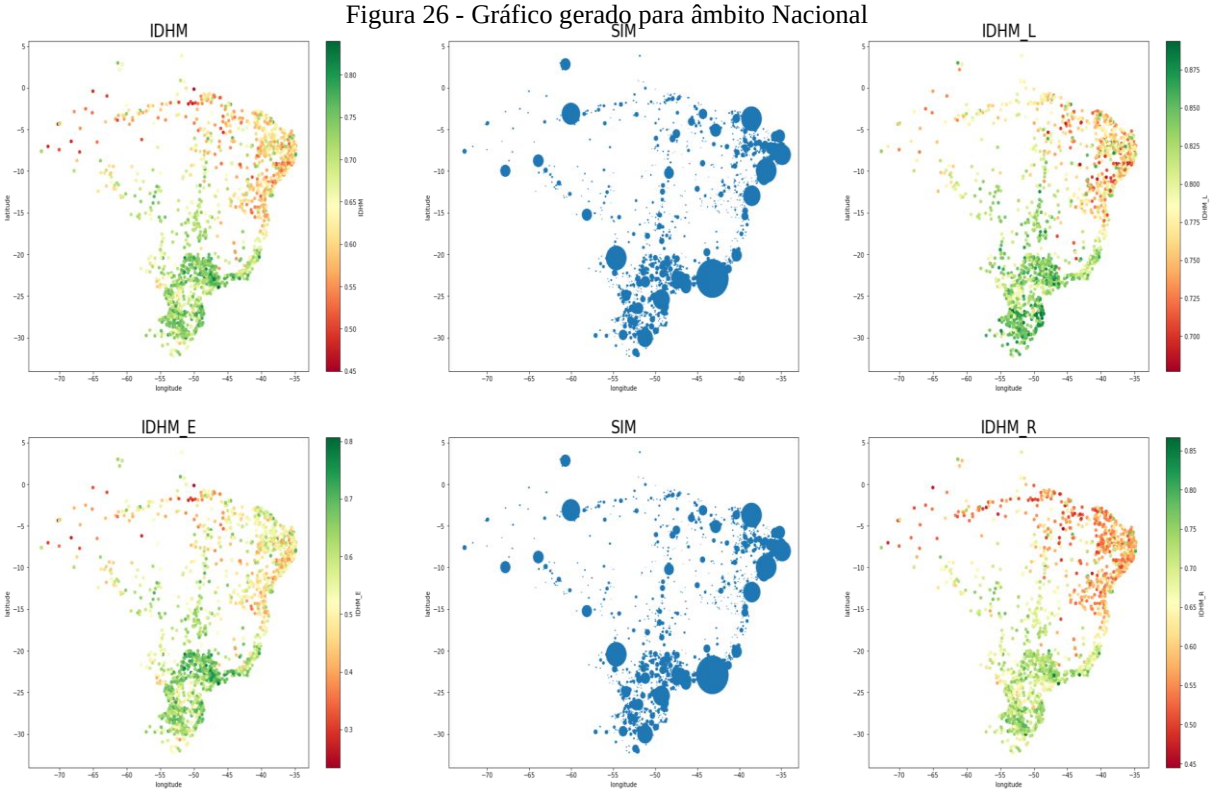
```
dfSragUF = df[df['SIGLA_UF'] == 'UF']
dfSragUF = gpd.GeoDataFrame(dfSragUF)
fig, ax = plt.subplots(2,3,figsize=(38,20))

ax[0,0].set_title('IDHM', color='black', size=26)
ax[0,1].set_title('SIM', color='black', size=26)
ax[0,2].set_title('IDHM_L', color='black', size=26)
ax[1,0].set_title('IDHM_E', color='black', size=26)
ax[1,1].set_title('SIM', color='black', size=26)
ax[1,2].set_title('IDHM_R', color='black', size=26)

dfSragUF.plot(column='IDHM',cmap='RdYlGn',edgecolor='black',legend=True,alpha=.5,ax=ax[0,0])
dfSragUF.plot(column='ID_MUNICIP',cmap='RdYlGn',edgecolor='black',legend=True,alpha=.5,ax=ax[0,1])
dfSragUF.plot(column='IDHM_L',cmap='RdYlGn',edgecolor='black',legend=True,alpha=.5,ax=ax[0,2])
dfSragUF.plot(column='IDHM_E',cmap='RdYlGn',edgecolor='black',legend=True,alpha=.5,ax=ax[1,0])
dfSragUF.plot(column='ID_MUNICIP',cmap='RdYlGn',edgecolor='black',legend=True,alpha=.5,ax=ax[1,1])
dfSragUF.plot(column='IDHM_R',cmap='RdYlGn',edgecolor='black',legend=True,alpha=.5,ax=ax[1,2])
```

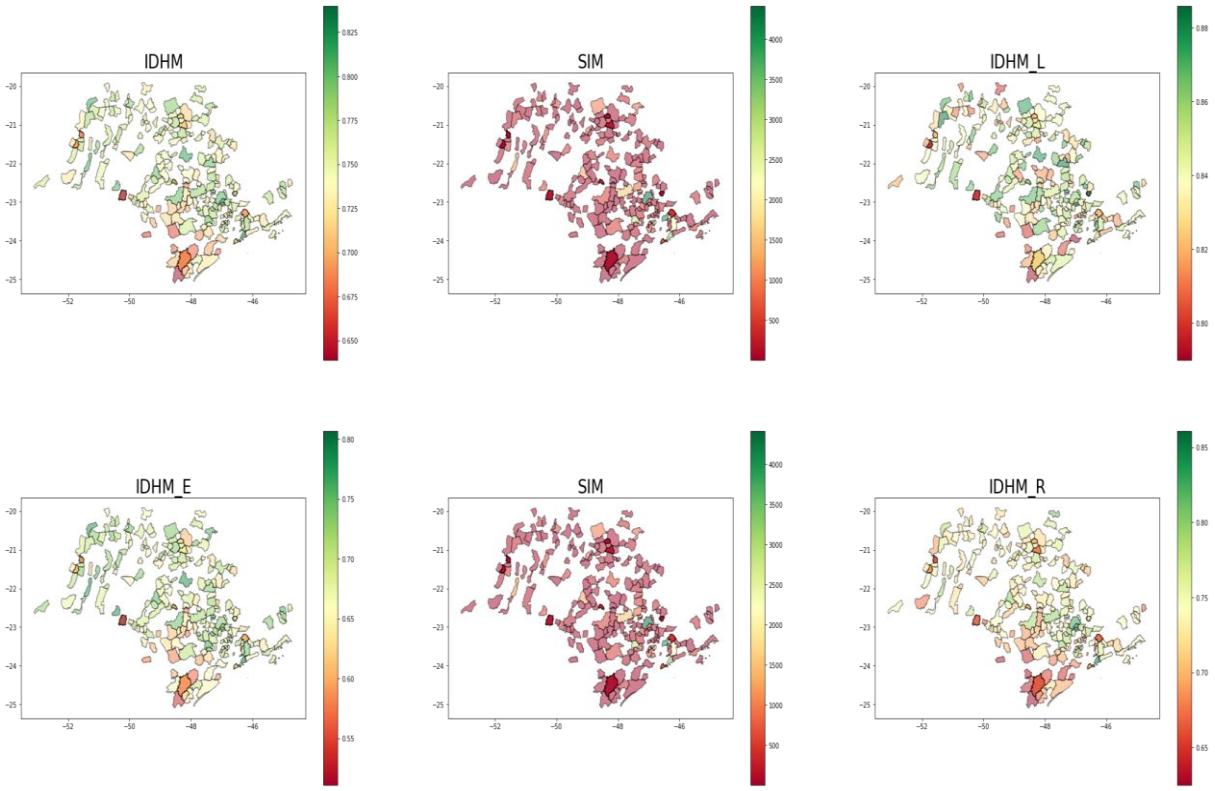
Fonte: Elaborado pelo autor.

Figura 26 - Gráfico gerado para âmbito Nacional



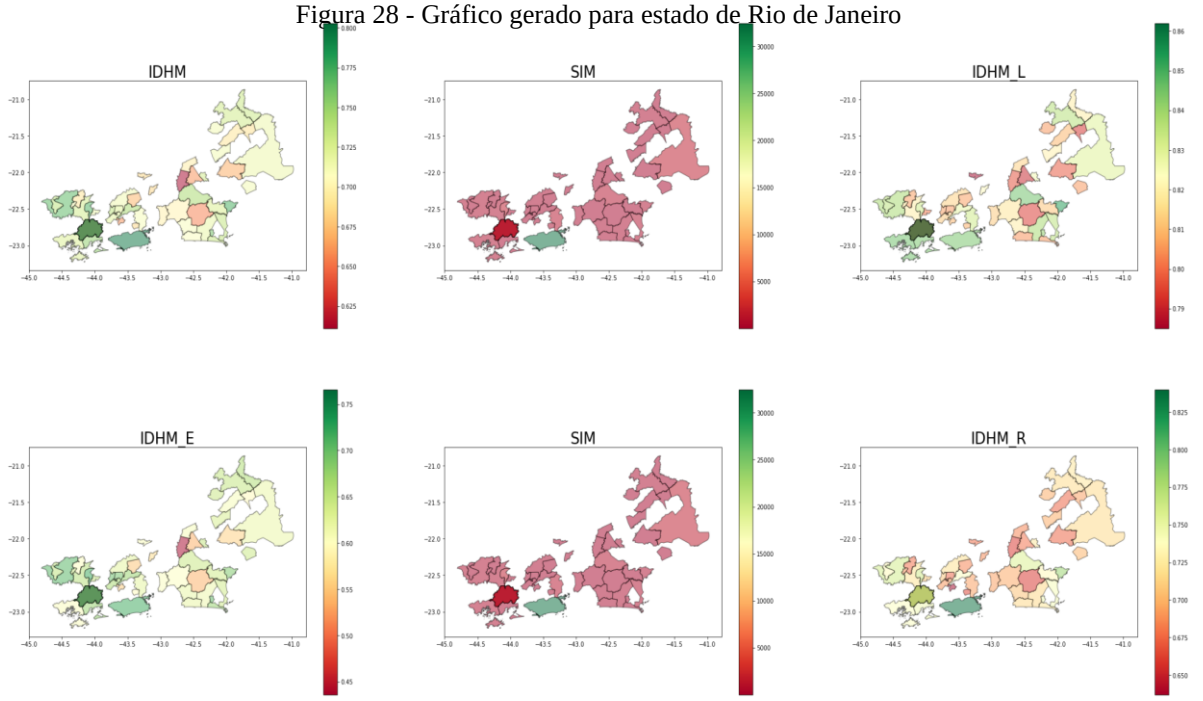
Fonte: Elaborado pelo autor.

Figura 27 - Gráfico gerado para o Estado de São Paulo



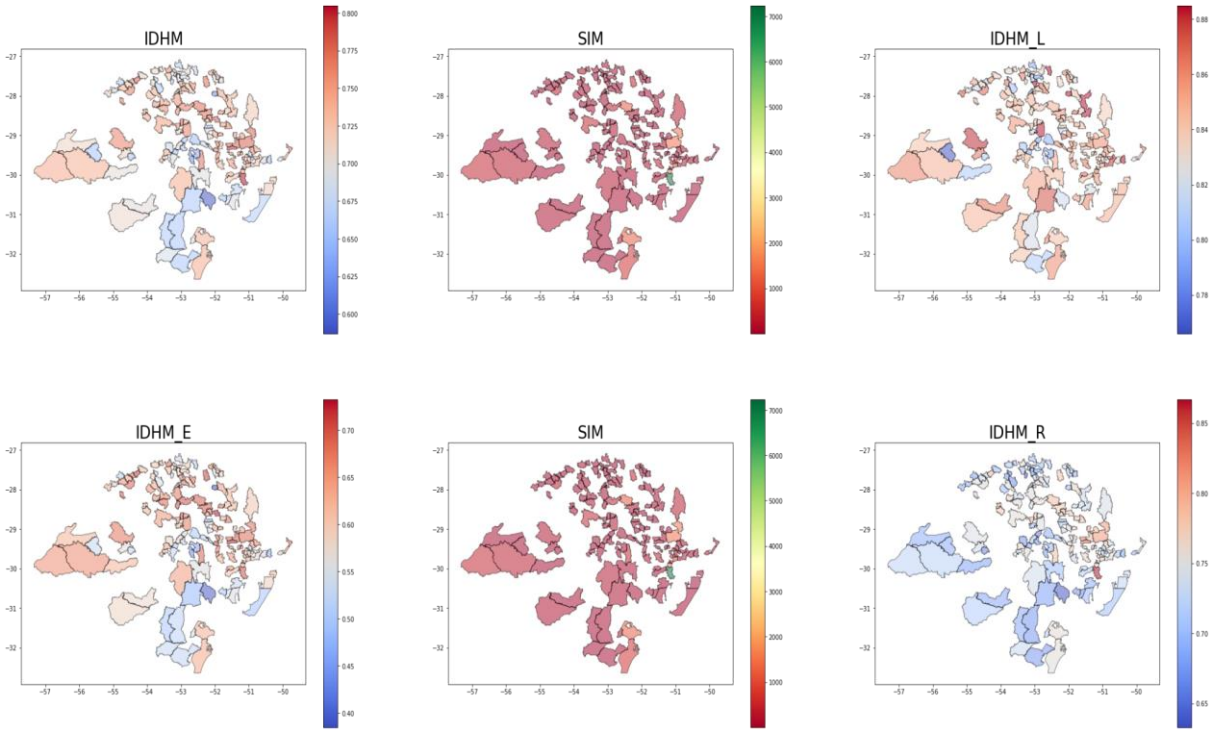
Fonte: Elaborado pelo autor.

Figura 28 - Gráfico gerado para estado de Rio de Janeiro



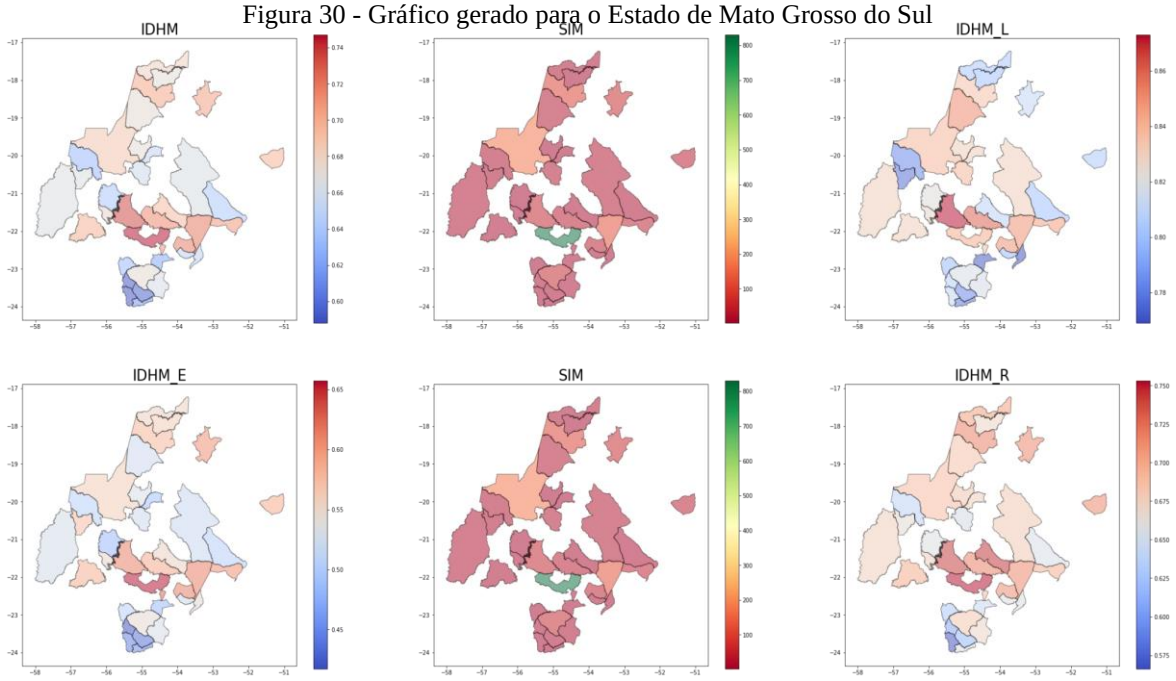
Fonte: Elaborado pelo autor.

Figura 29 - Gráfico gerado para O Estado de Rio Grande do Sul



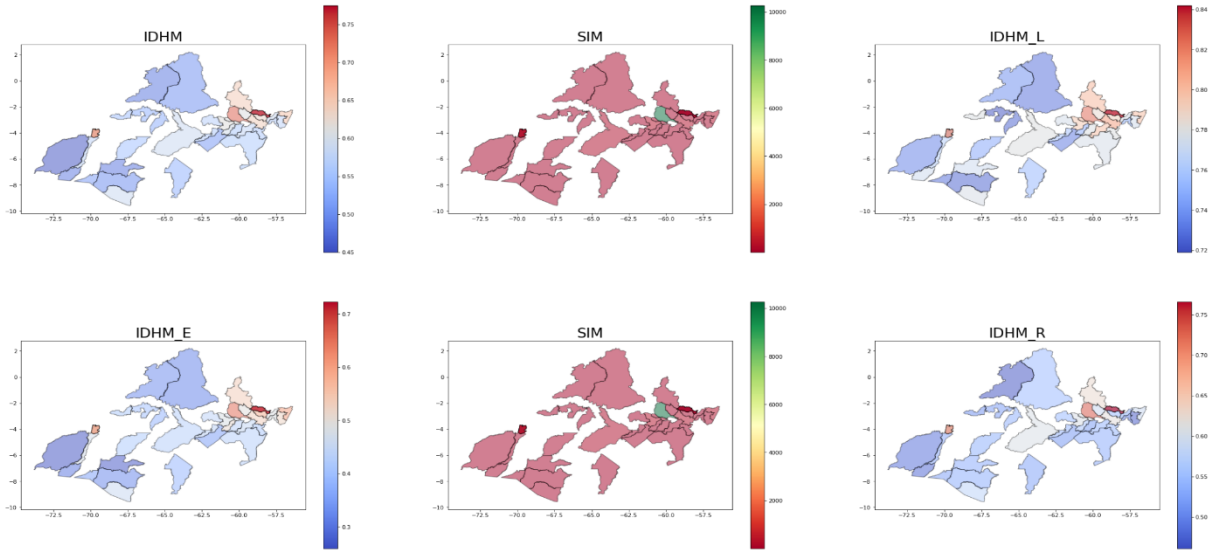
Fonte: Elaborado pelo autor.

Figura 30 - Gráfico gerado para o Estado de Mato Grosso do Sul



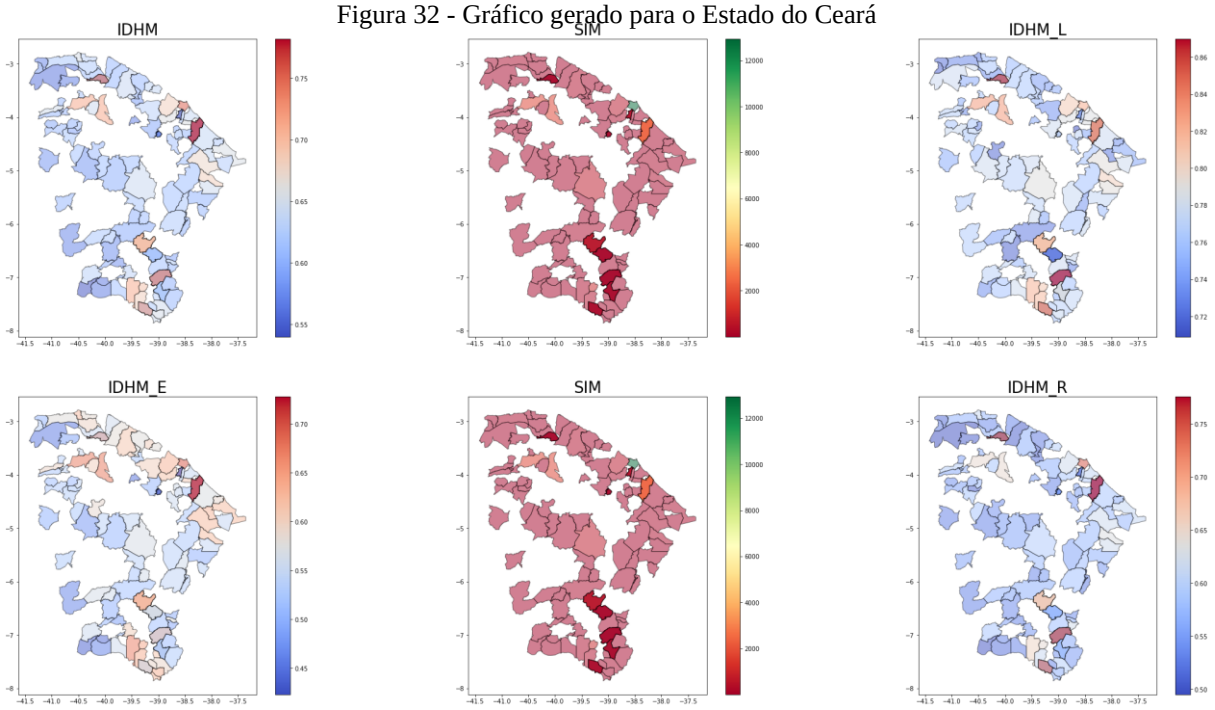
Fonte: Elaborado pelo autor

Figura 31 - Gráfico gerado para o Estado de Amazonas



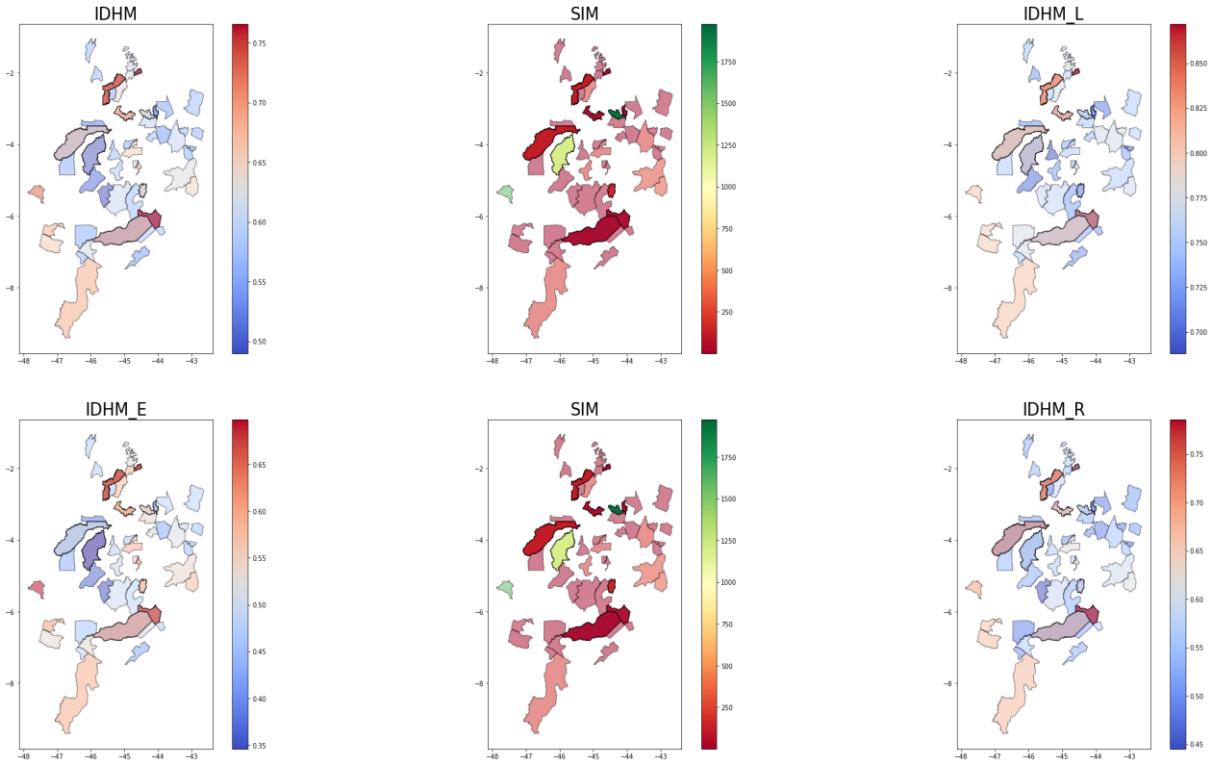
Fonte: Elaborado pelo autor

Figura 32 - Gráfico gerado para o Estado do Ceará



Fonte: Elaborado pelo autor

Figura 33 - Gráfico gerado para o Estado do Maranhão



Fonte: Elaborado pelo autor.

6 Conclusão

Ao fim desse trabalho, quando foram analisados separadamente todos os Estados brasileiros e os índices de mortes nacional, não foi possível inferir que os indicadores socioeconômicos adotados neste trabalho foram os fatores determinantes para o aumento das infecções ou da dispersão da COVID-19 durante a pandemia no ano de 2020 e do primeiro semestre de 2021.

Em algumas regiões do Sul e Centro foi possível visualizar certas áreas com indicadores maiores de IDH com menores índices de mortes e em outras regiões do Norte e Nordeste foi possível visualizar áreas com IDH menores e índice de morte maior, porém não é possível aplicar essas inferências para o ambiente nacional.

Sendo assim, conclui-se que, no contexto pandêmico vivido pelo Brasil, analisar apenas dados socioeconômicos não foram suficientes para determinar causa maior ou menor de mortes por COVID-19.

Perspectivas de Trabalhos Futuros

Em continuidade ao trabalho até aqui desenvolvido sugere-se realizar uma análise dos mapas obtidos e correlacionar os mesmos indicadores utilizados com microrregiões e regiões metropolitanas de grandes cidades para descobrir tendências e/ou variáveis importantes no processo.

A grande São Paulo, por exemplo através da Fundação Sistema Estadual de Análise de Dados (SEADE) disponibiliza os dados coletados, incluindo os indicadores de IDH das regiões, porém, para analisar tais dados será necessária uma quantidade bem maior de armazenamento e, consequentemente uma capacidade maior de processamento e análise.

Também propõe-se criar uma *Interface de Programação de Aplicações ou Interface de Programação de Aplicação* (API) adequada para realizar diretamente a busca, a seleção e a leitura dos dados na sua fonte de origem, facilitando-se assim a escolha dos Estados e cidades que deverão compor a análise e a visualização dos resultados graficamente.

Referências

SALLES, Silvana. **Ciência USP #28: A pandemia da ciência de dados**. [*S. l.*], 2020. Disponível em: <https://jornal.usp.br/podcast/ciencia-usp-28-a-pandemia-da-ciencia-de-dados/>. Acesso em: 4 jul. 2021.

COELHO, Lucas. Ciência de Dados: O que é, conceito e definição. *In*: **Ciência de Dados: O que é, conceito e definição**. [*S. l.*], 31 jul. 2020. Disponível em: <https://www.cetax.com.br/blog/data-science-ou-ciencia-de-dados>. Acesso em: 27 jun. 2021.

SILVEIRA, Debora Pricila. **O que é Data Science?**: Saiba mais sobre esta área que mistura Big Data, estatísticas e inteligência artificial e que vem ganhando destaque no Brasil, bem como em outros países.. [*S. l.*], 7 jul. 2016. Disponível em:<https://www.oficinadanet.com.br/post/16919-o-que-e-data-science>. Acesso em: 30 maio 2021.

ESCOLADEDADOS.ORG (ed.). **Mas o que significa isso?: Introdução à análise de dados**. [*S. l.*], 2019. Disponível em: <https://escoladedados.org/tutoriais/mas-o-que-significa-isso-introducao-a-analise-de-dados>. Acesso em: 30 jun. 2021.

ROVEDA, Ugo. **Google Colab: o que é, como usar e quais são as vantagens**?. [*S. l.*], 2019. Disponível em: <https://kenzie.com.br/blog/google-colab>. Acesso em: 1 jul. 2021.

MELO, Diego. **O que é Python? [Guia para iniciantes].** [*S. l.*], 26 jan. 2021. Disponível em: <https://tecnoblog.net/405640/o-que-e-python-guia-para-iniciantes>. Acesso em: 1 jul. 2021.

MATOS, David. **Por que Cientistas de Dados escolhem Python?**. [*S. l.*], 28 abr. 2019. Disponível em: <https://www.cienciaedados.com/por-que-cientistas-de-dados-escolhem-python>. Acesso em: 1 jul. 2021.