ANÁLISE EXPLORATÓRIA COM *PYTHON* EM BASE DE DADOS DE ÓBITOS NO BRASIL NO PERÍODO ENTRE 2019 E 2022

Krisna de Aquino Lira[1]; João Carlos Alchieri[2]

Resumo
Este estudo trata-se de uma análise exploratória da base de dados de óbitos do Brasil, disponibilizados pelo ministério da saúde no período entre 2019 e 2022, com objetivo de comparar e identificar sobre as principais causas de óbitos. A metodologia adotada foi de abordagem quantitativa, de natureza aplicada, com pesquisa exploratória, explicativa e de campo (coleta de dados), onde foram empregadas técnicas de estatística e bibliotecas em *Python* para análises dos dados. A relevância está na contribuição para o entendimento dos óbitos por Covid-19, identificação de tendências e padrões e na importância da análise dos dados oficiais para combater desinformações e notícias falsas relacionadas à pandemia. Os conjuntos de dados foram organizados por ano devido às fatalidades, e revelaram que o ano de 2021 se destacou como o mais letal da pandemia de Covid-19, com um total de 424.430 óbitos por Covid-19 registrados. Por conseguinte, optou-se por analisar detalhadamente os atributos relacionados a este ano, o que possibilitou identificar que a idade média dos falecidos foi de 64 anos, com um desvio padrão de 15,83. A análise dos dados também revelou que, em 2019, antes da pandemia, as doenças cardiovasculares já eram a principal causa de morte no Brasil. A análise exploratória de dados de óbitos por Covid-19 é uma ferramenta crucial para extrair informações, identificar padrões e tomar decisões informadas no combate à pandemia.
Palavras-chave: Análise exploratória; covid-19; python**.**

Abstract
This is an exploratory analysis of the Brazilian death database, made available by the Ministry of Health in the period between 2019 and 2022, with the objective of comparing and identifying the main causes of deaths. The methodology adopted was a quantitative approach, applied in nature, with exploratory, explanatory and field research (data collection), where statistical techniques and Python libraries were used for data analysis. The relevance lies in contributing to the understanding of Covid-19 deaths, identifying trends and patterns and the importance of analyzing official data to combat misinformation and fake news related to the pandemic. The datasets were organized by year by fatalities, and revealed that 2021 stood out as the deadliest of the Covid-19 pandemic, with a total of 424,430 Covid-19 deaths recorded. Therefore, it was decided to analyze in detail the attributes related to this year, which made it possible to identify that the average age of the deceased was 64 years old, with a standard deviation of 15.83. Data analysis also revealed that, in 2019, before the pandemic, cardiovascular diseases were already the main cause of death in Brazil. Exploratory analysis of Covid-19 death data is a crucial tool for extracting information, identifying patterns and making informed decisions in the fight against the pandemic.
Keywords: Exploratory analysis; covid-19; python**.**

[1] Mestra em Ciência da Computação pela Universidade Federal do Rio Grande do Norte-UFRN. E-mail: krisnaaquino@hotmail.com.
[2] Doutor em psicologia do desenvolvimento pela Universidade Federal do Rio Grande do Sul-UFRGS, professor titular do departamento de psicologia e orientador do Programa de Pós-Graduação em Ciência e Tecnologia e Inovação da Universidade Federal do Rio Grande do Norte-UFRN. E-mail: jcalchieri@gmail.com.

1 Introdução

Em virtude do volume e diversidade de informações produzidas diariamente, novas tecnologias e soluções vêm redefinindo a forma de se trabalhar com dados e estatísticas. A mineração de dados oferece estratégias para a análise dos dados em todas as áreas, sobretudo para o progresso e gerenciamento na área de saúde, seja encontrando padrões e correlações entre os dados, seja ajudando na tomada de decisões e desenvolvendo estratégias no gerenciamento das informações.

Segundo dados do Ministério da Saúde as Doenças Crônicas Não Transmissíveis (DCNT), incluindo doenças cardiovasculares, câncer, doenças pulmonares crônicas e diabetes, são responsáveis por três em cada cinco mortes em todo o mundo (BRASIL, 2021). No Brasil, o acesso inadequado a serviços de saúde de qualidade, que englobam a prevenção clínica, diagnósticos e o acesso a medicamentos essenciais, contribui para o aumento dessas doenças crônicas. Embora muitas mortes relacionadas a DCNT sejam evitáveis, a prevenção requer uma abordagem multidisciplinar e um planejamento abrangente. A atenção primária aliada à tecnologia desempenha um papel crucial na prevenção dessas doenças, concentrando-se na redução dos fatores de risco. A hipertensão, considerada uma das principais comorbidades agravantes da Covid-19, afeta um a cada quatro adultos no Brasil. As doenças cardiovasculares (infarto, hipertensão, AVC e outras enfermidades) foram responsáveis por uma média de 34 óbitos por hora, 829 óbitos por dia e mais de 302 mil óbitos no ano de 2017. Esse é o retrato das doenças cardiovasculares no Brasil, que têm como principal fator de risco a hipertensão arterial, conforme dados do Sistema de Informações de Mortalidade (SIM), do Ministério da Saúde (BRASIL, 2019). Mesmo sendo uma doença de diagnóstico relativamente simples, metade dos afetados não têm conhecimento de sua condição, e aproximadamente 35% da população brasileira lida com essa comorbidade.

O estudo tem como objetivo comparar e identificar as principais causas de óbitos no período de 2019 a 2022, por meio das bases de dados acessíveis no site DATASUS (OPEN DATASUS) e realizar uma análise exploratória dos dados com o intuito de identificar padrões relevantes nos óbitos por Covid-19. A análise da base de dados referente ao ano 2019 foi utilizada para comparar as causas de óbitos no Brasil antes do surgimento da pandemia de Covid-19.

2 Referencial Teórico

Este capítulo é dedicado à base teórica do tópico sob análise e está dividido em cinco seções. São abordados conceitos de Big Data, tipos de dados, mineração de dados, pré-processamentos dos dados e da linguagem de programação utilizada no desenvolvimento deste trabalho, *Python*.

2.1 Big Data

Conforme enfatiza Machado (2018), o mundo contemporâneo está experimentando um aumento exponencial na quantidade de informações disponíveis. Esse crescimento não é apenas resultado das atividades empresariais, mas também se deve à proliferação da internet, das redes sociais, dos smartphones e dos dispositivos móveis. Esses avanços tecnológicos estão contribuindo para a criação de grandes volumes de dados, que podem variar em complexidade e abranger informações estruturadas e não estruturadas.

Machado (2018) menciona que até o início do século XXI, cerca de 25% de todas as informações do mundo eram digitais, com uma quantidade significativa de dados ainda existindo em formatos físicos, como documentos impressos e livros. No entanto, entre os anos de 2012 e 2014, ocorreu uma mudança dramática, e aproximadamente 98% de todas as informações passaram a existir em formato digital. Esse aumento foi possibilitado pela redução dos custos de computadores e sistemas de armazenamento de dados, bem como pelo crescimento exponencial das capacidades de processamento.

A disseminação das informações digitais tornou-se uma realidade, e os dados passaram a desempenhar um papel de extrema importância. Isso se deve ao fato de que os dados podem proporcionar insights valiosos através da identificação de correlações. Esses insights podem abranger desde o comportamento de compra de clientes individuais até a capacidade de prever crises em setores econômicos, entender a migração de clientes entre empresas e detectar surtos de doenças infecciosas, como a gripe H1N1 ou o Zika Vírus.

De acordo com Ribeiro Neto (2020), o termo Big Data surgiu para se referir às aplicações de computadores que utilizam grandes volumes de dados em diferentes formatos, que podem ser agrupados, lidos, convertidos, analisados com técnicas estatísticas, matemáticas e computacionais, gerando um novo tipo de conhecimento chamado de “Data Insight”, algo conclusivo e ainda nunca pensado sobre os dados originais. A geração de insights a partir dos dados, pode resultar tanto na decisão de uma mudança brusca na orientação de um negócio,

quanto na criação de um novo produto, chamado de “Data-Driven Product” (Produto Orientado a Dados) que podem revolucionar a empresa e os seus negócios.

Conforme mencionado por Ribeiro Neto (2020), os dados destinados ao Big Data originam-se de três principais fontes: indivíduos, máquinas e empresas.

1 - Pessoas geram dados a partir de Redes Sociais (Facebook, Twitter, Instagram, LinkedIn), e-mails, uso de Internet, geração de documentos, publicação de blogs, etc.

2 - Máquinas geram dados a partir de sensores, satélites, arquivos de logs, câmeras, máquinas de sequenciamento genético, telescópios espaciais, sondas, etc.

3 - Empresas geram dados a partir de transações comerciais, cartões de crédito, sistemas de controles administrativos e financeiros, comércio eletrônico, registros médicos, vendas de produtos, pesquisa de novas tecnologias, etc.

2.2 Tipos de Dados

Conforme Hernández e Mendoza (2018, p. 224), dados representam a matéria-prima essencial para análise, constituindo as bases fundamentais para a construção do conhecimento. Os referidos autores também delineiam os seguintes procedimentos para a coleta de dados quantitativos:

Definir a forma ideal de coletar os dados quantitativos de acordo com a declaração do problema e método implementado (escopo, hipótese, design e amostra).

Escolher ou desenvolver um ou mais instrumentos ou métodos para coletar os dados necessários.

Aplicar os instrumentos ou métodos (medir as variáveis nos casos).

Obter os dados.

Codificar os dados.

Arquivar dados e prepará-los para análise estatística computacional.

Por sua vez, McKinney (2023) descreve que a análise de dados faz uso de dados estruturados, um conceito propositadamente amplo que engloba várias formas comuns de dados, incluindo:

Dados tabulares ou semelhantes a planilhas, onde cada coluna pode ser de um tipo diferente (string, numérico, data ou outro). Isso engloba a maioria dos tipos de dados geralmente armazenados em bancos de dados relacionais ou em arquivos de texto delimitados por tabulações ou vírgulas.

Matrizes multidimensionais.

Múltiplas tabelas de dados interconectadas por colunas-chave (que podem ser chaves primárias ou estrangeiras para um usuário de SQL).

Séries temporais com espaçamento uniforme ou desigual.

2.3 Mineração de dados

De acordo com Tan et al. (2019), a mineração de dados é a ação de identificar informações valiosas, padrões, tendências e conhecimentos ocultos em extensos conjuntos de dados. Envolve a análise de dados brutos para encontrar relações, correlações ou padrões significativos que possam ser usados para tomar decisões bem fundamentadas.

Conforme destacado por Fawcett e Provost (2018), devido à ampla disponibilidade de volumes de dados, empresas em quase todos os setores estão direcionando seus esforços para explorar esses recursos em busca de vantagens competitivas. No passado, as empresas contavam com equipes de estatísticos e analistas para realizar análises manuais de conjuntos de dados. No entanto, a escala e a diversidade desses conjuntos de dados ultrapassaram significativamente a capacidade da análise manual.

Simultaneamente, o avanço considerável na capacidade de processamento computacional, a disseminação da comunicação em rede e o desenvolvimento de algoritmos permitiram a conexão de diversos conjuntos de dados, possibilitando análises muito mais abrangentes e aprofundadas do que eram viáveis anteriormente. A convergência desses fatores deu origem à crescente adoção de princípios da ciência de dados e técnicas de mineração de dados no mundo empresarial.

Uma das aplicações mais proeminentes de técnicas de mineração de dados se encontra no campo do marketing, sendo utilizada em iniciativas como marketing direcionado, publicidade online e recomendações para vendas cruzadas. Na gestão de relacionamento com o cliente, a mineração de dados é empregada para analisar o comportamento dos clientes, com o objetivo de aprimorar a retenção e otimizar o valor esperado deles.

No setor financeiro, a mineração de dados desempenha um papel crucial em atividades como classificação e negociação de crédito, além de ser fundamental na detecção de fraudes e na administração da força de trabalho. Os autores citados afirmam que a mineração de dados é uma arte, pois requer a aplicação de uma quantidade substancial de ciência e tecnologia, mas sua aplicação adequada também envolve uma dimensão artística. No entanto, assim como muitas artes maduras, existe um processo bem definido que oferece uma estrutura para abordar o problema, garantindo consistência, repetibilidade e objetividade razoável.

2.4 Pré-processamento de Dados

O pré-processamento de dados como uma etapa fundamental no processo de mineração de dados, que envolve a preparação dos dados antes de aplicar técnicas analíticas. Essa fase desempenha um papel crucial na garantia de que os resultados da mineração sejam precisos, confiáveis e relevantes. (Tan et al., 2019). O processo de limpeza de dados envolve tarefas como remover dados ausentes, substituir dados ausentes ou inconsistentes, aplicar transformações e conversões em dados numéricos, ou outros tipos de dados e normalizar dados. O objetivo da limpeza de dados é tornar os dados mais confiáveis e consistentes para que possam ser usados para análise, ou seja com mais qualidade.

As principais técnicas do pré-processamento de dados são:

Agregação

Amostragem

Redução de dimensionalidade

Seleção de subconjuntos de recursos

Criação de recursos

Discretização e binarização

Transformação de variáveis

Gaspar *et al*. (2023) destaca que este processo demanda um tempo importante em toda a análise de dados e ele recomenda destinar de 20% a 40% do tempo do projeto de análise para esta etapa.

2.5 Análise de dados com *Python*

O *Python* é uma linguagem de programação de alto nível e de propósito geral, criada com o objetivo de ser uma linguagem fácil para implementação de algoritmos, possuindo uma sintaxe simples e intuitiva (Netto A, 2021).

Ao desenvolver aplicações científicas, é aconselhável utilizar a distribuição anaconda, uma vez que é de código aberto e representa a maneira mais simples de executar aplicações científicas desenvolvidas em *Python*, compatível com Linux, Windows e Mac OS X. Dentro da distribuição Anaconda, encontra-se o Jupyter Notebook, uma interface bastante útil para criar modelos e compartilhá-los facilmente.

Uma alternativa semelhante ao Jupyter Notebook, que não exige configuração prévia, é uma ferramenta desenvolvida pelo Google chamada Colaboratory. Para utilizar esse ambiente, é necessário apenas possuir uma conta Gmail, pois todos os notebooks serão armazenados no Google Drive.

As bibliotecas na linguagem *Python* utilizadas para o processamento dos dados utilizados neste trabalho são:

• Pandas: A biblioteca Pandas tem como objetivo proporcionar ao *Python* um conjunto de ferramentas integradas para o processamento de dados estruturados que estão organizados em forma de tabelas ou bancos de dados, onde esses dados são dispostos em linhas e colunas, com uma indexação padronizada (Feltrin, 2021).

• Seaborn: É uma biblioteca de visualização de dados *Python* baseada no matplotlib. Ela fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos (SEABORN, 2023).

• Matplotlib: O Matplotlib é uma biblioteca de plotagem 2D do *Python* que produz números de qualidade de publicação em vários formatos de cópia impressa e ambientes interativos entre plataformas. O Matplotlib pode ser usado em scripts *Python*, nos shell *Python* e *Python*, notebooks Jupyter e em servidores web. (MATPLOTIB, 2023).

• DataFrames: É um tipo de dado básico usado na biblioteca Pandas, semelhante a uma série (inclusive um DataFrame é formado por um conjunto de séries), porém equivalente a uma array multidimensional (Feltrin, 2021).

3 Método

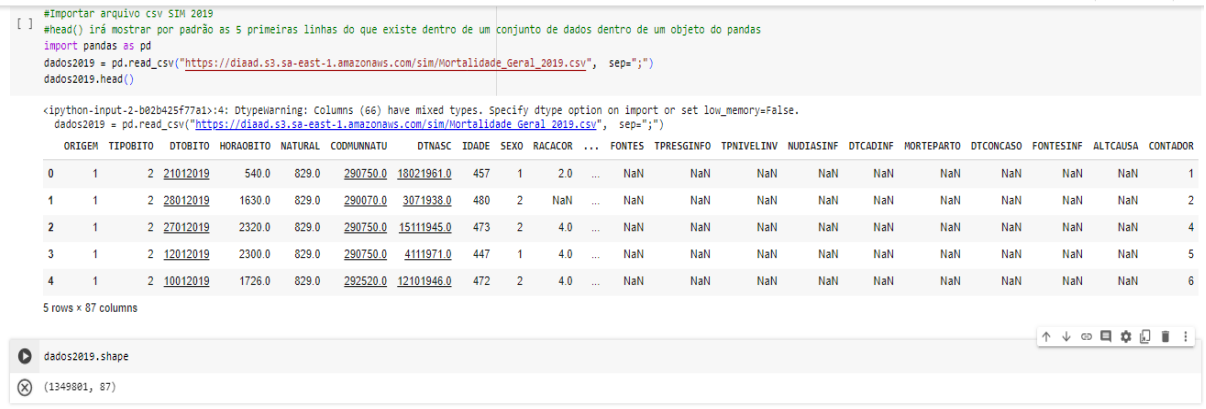
A metodologia adotada seguiu uma abordagem quantitativa e de natureza aplicada, com uma pesquisa exploratória, explicativa e de campo, na qual foram utilizadas técnicas estatísticas e bibliotecas em Python para a análise dos dados coletados.

Para o estudo e desenvolvimento deste projeto, foram empregados dados, provenientes do DATASUS (OPEN DATASUS), disponibilizado pelo Ministério da Saúde, que é um sistema de vigilância epidemiológica nacional, cujo objetivo é captar dados sobre os óbitos do país a fim de fornecer informações sobre mortalidade para todas as instâncias do sistema de saúde. A análise dessas informações permite estudos não apenas do ponto de vista estatístico e epidemiológico, mas também sócio-demográfico. O documento base para a captação dos dados de mortalidade é a Declaração de Óbito (DO). No primeiro momento as bases foram importadas e em seguida foi iniciada a etapa de pré-processamento, utilizando-se da linguagem de programação *Python* através do *google colab*, ou *Colaboratory*, é uma plataforma gratuita baseada na nuvem oferecida pelo *Google* que permite criar, compartilhar, e executar *notebooks* interativos de *Python,* auxiliado pelas bibliotecas Pandas e *Numpy.*

Foram importados e analisados dados de mortalidade geral no Brasil no período entre 2019 e 2022, mas definido o ano de 2021 para o detalhamento e dos resultados, o ano mais

letal da pandemia de Covid-19 (World Health Organization). Após importação do arquivo de óbitos Brasil no ano de 2019, foi feito conhecimento e análise da base de dados através do *Python*, aplicando o comando *shape* ao *dataset*, pode-se verificar a quantidade de colunas/atributos foi igual à 87 colunas e 1.349.801 linhas (Figura 1), no arquivo do ano de 2020 conteve 87 colunas e 1.556.824 linhas (Figura 2), já no arquivo de óbitos do ano de 2021, pode se verificar que o *dataset* contém 87 colunas e 1.832.649 linhas (Figura 3) e no arquivo de óbitos do ano de 2022 o *dataset* conteve 87 colunas 1.542.158 linhas (Figura 4).

Figura 1 - Importação dados óbitos 2019



```
[ ] #Importar arquivo csv SIM 2019
    #head() irá mostrar por padrão as 5 primeiras linhas do que existe dentro de um conjunto de dados dentro de um objeto do pandas
    import pandas as pd
    dados2019 = pd.read_csv("https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral_2019.csv", sep=";")
    dados2019.head()
```

```
<ipython-input-2-b02b425f77a1>:4: DtypeWarning: Columns (66) have mixed types. Specify dtype option on import or set low_memory=False.
  dados2019 = pd.read_csv("https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral_2019.csv", sep=";")
```

	ORIGEM	TIPOBITO	DTOBITO	HORAOBITO	NATURAL	CODMUNNATU	DTNASC	IDADE	SEXO	RACACOR	...	FONTES	TPRESGINFO	TPNIVELINV	NUDIASINF	DTCADINF	MORTEPARTO	DTCONCASO	FONTESINF	ALTCAUSA	CONTADOR
0	1	2	21012019	540.0	829.0	290750.0	18021961.0	457	1	2.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
1	1	2	28012019	1630.0	829.0	290070.0	3071938.0	480	2	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
2	1	2	27012019	2320.0	829.0	290750.0	15111945.0	473	2	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4
3	1	2	12012019	2300.0	829.0	290750.0	4111971.0	447	1	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5
4	1	2	10012019	1726.0	829.0	292520.0	12101946.0	472	2	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6

5 rows × 87 columns

```
dados2019.shape
```

```
(1349801, 87)
```

Fonte: Autoria própria (2023).

Figura 2 - Importação dados óbitos 2020 google colab

```
[ ] #Importar arquivo csv SIM 2020
    #head() irá mostrar por padrão as 5 primeiras linhas do que existe dentro de um conjunto de dados dentro de um objeto do pandas
    import pandas as pd
    dados2020 = pd.read_csv("https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral_2020.csv", sep=";")
    dados2020.head()
```

```
<ipython-input-2-bad2290493a9>:4: DtypeWarning: Columns (66,79) have mixed types. Specify dtype option on import or set low_memor
  dados2020 = pd.read_csv("https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral_2020.csv", sep=";")
```

	ORIGEM	TIPOBITO	DTOBITO	HORAOBITO	NATURAL	CODMUNNATU	DTNASC	IDADE	SEXO	RACACOR	...	FONTES	TPRESGINFO	TPNIVELIN
0	1	2	18052020	1452.0	812.0	120010.0	7061932.0	487	1	4.0	...	NaN	NaN	Nal
1	1	2	20052020	2115.0	812.0	120010.0	16031952.0	468	2	4.0	...	NaN	NaN	Nal
2	1	2	21052020	1200.0	812.0	120010.0	17021961.0	459	2	4.0	...	NaN	NaN	Nal
3	1	2	21052020	1233.0	812.0	120010.0	10081942.0	477	2	4.0	...	NaN	NaN	Nal
4	1	2	22052020	730.0	812.0	120030.0	13041936.0	484	2	4.0	...	NaN	NaN	Nal

5 rows × 87 columns

```
[ ] #retorna uma tupla que representa linhas e colunas como o valor das dimensões
    dados2020.shape
```

```
(1556824, 87)
```

Fonte: Autoria própria (2023).

Figura 3 - Importação dados óbitos 2021 google colab

```
#Importar arquivo csv SIM 2021
#head() irá mostrar por padrão as 5 primeiras linhas do que existe dentro de um conjunto de dados dentro de um objeto do pandas
import pandas as pd
#(importando arquivo csv completo)
dados2021 = pd.read_csv("https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIM/Mortalidade_Geral_2021.csv",  sep=";")
dados2021.head()
```

```
<ipython-input-3-e5ecfd33e855>:5: DtypeWarning: Columns (66) have mixed types. Specify dtype option on import or set low_memory=
  dados2021 = pd.read_csv("https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIM/Mortalidade_Geral_2021.csv",  sep=";")
```

	ORIGEM	TIPOBITO	DTOBITO	HORAOBITO	NATURAL	CODMUNNATU	DTNASC	IDADE	SEXO	RACACOR	...	FONTES	TPRESGINFO	TPNIVEI
0	1	2	23032021	1500.0	811.0	110020.0	18061962.0	458	1	4.0	...	NaN	NaN	
1	1	2	23032021	243.0	812.0	120050.0	19021971.0	450	1	4.0	...	NaN	NaN	
2	1	2	23032021	1310.0	812.0	120040.0	1101956.0	464	2	4.0	...	NaN	NaN	
3	1	2	17042021	2149.0	812.0	120050.0	6011999.0	422	1	4.0	...	NaN	NaN	
4	1	2	6012021	420.0	812.0	120020.0	20082020.0	304	1	4.0	...	SXXSXX	NaN	

5 rows × 87 columns

```
#dataframe escolhido
dados2021.shape
```

```
(1832649, 87)
```

Fonte: Autoria própria (2023).

Figura 4 - Importação dados óbitos 2022 google colab

```
#Importar arquivo csv SIM 2022
#head() irá mostrar por padrão as 5 primeiras linhas do que existe dentro de um conjunto de dados dentro de um objeto do par
dados2022 = pd.read_csv("https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIM/DO22OPEN.csv",  sep=";")
dados2022.head()
```

```
<ipython-input-4-3ea95d66e0e6>:3: DtypeWarning: Columns (65) have mixed types. Specify dtype option on import or set low_mem
  dados2022 = pd.read_csv("https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIM/DO22OPEN.csv",  sep=";")
```

	contador	ORIGEM	TIPOBITO	DTOBITO	HORAOBITO	NATURAL	CODMUNNATU	DTNASC	IDADE	SEXO	...	TPRESGINFO	TPNIVELINV
0	1	1	2	9022022	1409.0	826.0	261160.0	21011972.0	450	2	...	NaN	NaN
1	2	1	2	1022022	NaN	NaN	NaN	NaN	999	1	...	NaN	NaN
2	3	1	2	9022022	1500.0	826.0	260260.0	26091932.0	489	2	...	NaN	NaN
3	4	1	2	2022022	2345.0	826.0	261160.0	6061968.0	453	1	...	NaN	NaN
4	5	1	2	11022022	1706.0	826.0	261640.0	25041930.0	491	2	...	NaN	NaN

5 rows × 86 columns

```
dados2022.shape
```

```
(1542158, 86)
```

Fonte: Autoria própria (2023).

Ainda nesta fase de pré-processamento dos dados, foram realizadas análises mais profundas da base (Gaspar *et al.* 2023). Após a compreensão dos dados, é crucial identificar dados ausentes, inconsistências ou possíveis erros. Essa inspeção pode envolver a visualização dos dados importados, a análise das primeiras e últimas linhas do arquivo, a verificação das estatísticas descritivas para valores numéricos, a revisão de formatos de datas, e a avaliação de campos de texto livre ou categóricos. No caso de campos de texto, é importante verificar a

presença de caracteres especiais não formatados ou espaços antes ou após os textos e de dados ausentes.

4 Resultados e Discussão

4.1 Resultados

As bases de dados foram categorizadas pela quantidade de ocorrências e pelas causas dos óbitos, através da Classificação Internacional de Doenças (CID) por ano. Conforme ilustrado na Figura 5, no *dataset* referente o ano 2019 observa-se em primeiro lugar, o CID I219 (Infarto agudo do miocárdio não especificado) com 95.689 registros de óbitos, em seguida o CID J 189 (Pneumonia não especificada) com 50.044 registros de óbitos e em terceiro lugar o CID R99 (Outras causas mal definidas e não especificadas de mortalidade) com 48.511 registros de óbitos.

Figura 5 - Causas de óbitos 2019

```
[ ] contagem_causa2019 = dados2019['CAUSABAS'].value_counts().reset_index()
    contagem_causa2019 = contagem_causa2019.rename(columns={'index': 'CAUSABAS', 'CAUSABAS': 'quantidade'})
    contagem_causa2019 = contagem_causa2019.sort_values(by='quantidade', ascending=False)
    print(contagem_causa2019)
```

```
     CAUSABAS  quantidade
0        I219       92689
1        J189       50044
2         R99       48511
3         I64       33895
4        E149       28613
...       ...         ...
4825     V835           1
4826     D433           1
4827     B719           1
4828     X320           1
5514     F984           1

[5515 rows x 2 columns]
```

Fonte: Autoria própria (2023).

Com início da pandemia de Covid-19 em 2020, como pode ser observado no *dataset* referente ao ano 020 na Figura 6 aparece o CID B342 (Infecção por coronavírus de localização não especificada) como principal causa de óbitos, com 212.706 registros de óbitos, seguido do CID I219 (Infarto agudo do miocárdio não especificado) com 87.961 registros de óbitos e em terceiro lugar o CID CID R99 (Outras causas mal definidas e não especificadas de mortalidade) com 58.899 registros de óbitos.

Figura 6 - Causas de óbitos 2020

```
[ ] contagem_causa2020 = dados2020['CAUSABAS'].value_counts().reset_index()
    contagem_causa2020 = contagem_causa2020.rename(columns={'index': 'CAUSABAS', 'CAUSABAS': 'quantidade'})
    contagem_causa2020 = contagem_causa2020.sort_values(by='quantidade', ascending=False)
    print(contagem_causa2020)
```

```
     CAUSABAS  quantidade
0        B342      212706
1        I219       87961
2         R99       58899
3         I10       37600
4        J189       36907
...       ...         ...
4842     R908           1
4843     Y443           1
4844     P743           1
4845     H652           1
5497     X792           1

[5498 rows x 2 columns]
```

Fonte: Autoria própria (2023).

O ano de 2021, o mais letal da pandemia de Covid-19, conforme pode ser observado na Figura 7, mostra o CID B342 (Infecção por coronavírus de localização não especificada) em primeiro lugar, com a quantidade alarmante de 424.430 registros de óbitos. Em seguida, o CID I219 (Infarto agudo do miocárdio não especificado) com 93.297 registros, seguido pelo CID R99 (Outras causas mal definidas e não especificadas de mortalidade) com 60.502 registros e o CID I10 (Hipertensão essencial primária) com 39.963 registros de óbitos.

Figura 7 - Causas de óbitos 2021

```
[ ] contagem_causa2021 = dados2021['CAUSABAS'].value_counts().reset_index()
    contagem_causa2021 = contagem_causa2021.rename(columns={'index': 'CAUSABAS', 'CAUSABAS': 'quantidade'})
    contagem_causa2021 = contagem_causa2021.sort_values(by='quantidade', ascending=False)
    print(contagem_causa2021)
```

```
     CAUSABAS  quantidade
0        B342      424461
1        I219       93348
2         R99       61098
3         I10       39966
4         I64       35808
...       ...         ...
4873     X448           1
4874     H652           1
4875     X815           1
4876     X251           1
5572     E012           1

[5573 rows x 2 columns]
```

Fonte: Autoria própria (2023).

Já no dataset referente o ano 2022, a principal causa de óbito voltou a ser o CID I219 (Infarto agudo do miocárdio não especificado), como antes da pandemia de Covid-19 com 95.072 registros de óbitos, em segundo lugar o CID B342 (Infecção por coronavírus de localização não especificada) com 65.764 registros e em terceiro lugar o CID R99 (Outras causas mal definidas e não especificadas de mortalidade) com 51.261 registros de óbitos.

Figura 8 - Causas de óbitos 2022

```
[ ]  contagem_causa2022 = dados2022['CAUSABAS'].value_counts().reset_index()
     contagem_causa2022 = contagem_causa2022.rename(columns={'index': 'CAUSABAS', 'CAUSABAS': 'quantidade'})
     contagem_causa2022 = contagem_causa2022.sort_values(by='quantidade', ascending=False)
     print(contagem_causa2022)
```

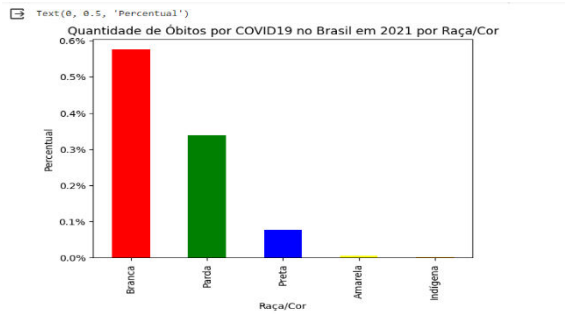
```
     CAUSABAS  quantidade
0        I219       95072
1        B342       65764
2         R99       51261
3        J189       48072
4         I10       39220
...       ...         ...
4884     K092           1
4885     O689           1
4886     M620           1
4887      R69           1
5538     P119           1

[5539 rows x 2 columns]
```

Fonte: Autoria própria (2023).

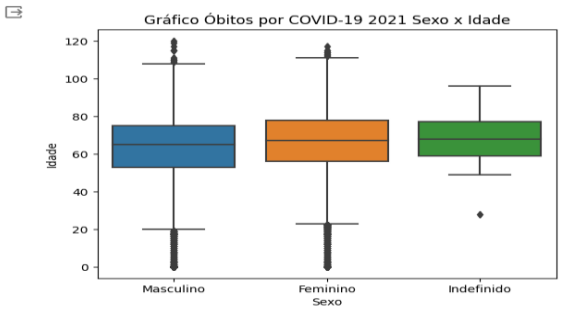
Foram gerados gráficos estatísticos e de visualização no arquivo de óbitos por Covid-19 do ano 2021 para a descoberta de padrões de forma rápida e intuitiva. Foram utilizados os atributos raça/cor (Figura 9), sexo e idade (Figura 10).

Figura 9 - Óbitos Covid-19 no Brasil 2021 por raça/cor



Fonte: (Autoria própria, 2023).

Figura 10 - Óbitos Covid-19 no Brasil 2021 sexo e idade



Fonte: (Autoria própria, 2023).

Após aplicado o filtro de Covid-19 ao *dataset* dados 2021, foram calculadas medidas de tendência para o atributo idade, conforme ilustrado na Figura 11. A média de idade dos óbitos por Covid-9 foi de 64 anos, a mediana 66 anos e a moda 0 (zero) e 66 anos.

Figura 11- Medidas de tendência óbitos Covid-19 no Brasil 2021

```
# Medidas de tendência
# Calcular as medidas de tendência central para idade  após filtro covid
# Média
x = dados2021.IDADE2.mean()
print('Média:', round(x, 2))
# Mediana
y = dados2021.IDADE2.median()
print('Mediana:', y)
# Moda
z = dados2021.IDADE2.mode()
print('Moda:', z)
```

```
Média: 64.83
Mediana: 66.0
Moda: 0    66
Name: IDADE2, dtype: int64
```

F onte: (Autoria própria, 2023).

Igualmente realizadas medidas de dispersão para a variável idade no arquivo de óbitos por Covid-19 no Brasil no ano de 2021 valor mínimo é 0 (zero) anos, o valor máximo é 120 anos, a variância é de 250.62 e o desvio padrão é de 15.83 (Figura 12).

Figura 12 - Medidas de dispersão óbitos Covid-19 no Brasil 2021

```
#Medidas de dispersão
# Calcular as medidas de dispersão da idade
# Mínimo
min = dados2021.IDADE2.min()
# Máximo
max = dados2021.IDADE2.max()
# Intervalo
print('Intervalo:', min, ' - ', max)
# Variância
var = dados2021.IDADE2.var()
print('Variância:', round(var, 2))
# Desvio padrão
std = dados2021.IDADE2.std()
print('Desvio Padrão:', round(std, 2))
```

```
Intervalo: 0  -  120
Variância: 250.62
Desvio Padrão: 15.83
```

Fonte: (Autoria própria, 2023).

A distribuição de frequência absoluta e relativa para variável sexo no arquivo de óbitos por Covid-19 no Brasil no ano de 2021 e os resultados estão representados na Figura 13.

Figura 13- Distribuição de frequência por sexo óbitos Covid-19 no Brasil 2021

```
[57] # Distribuição de frequência
     # Frequência absoluta
     fa = dados2021.SEXO.value_counts()
     print('Frequência absoluta')
     print(fa)
     # Frequência relativa
     fr = dados2021.SEXO.value_counts(normalize = True)
     print('Frequência relativa')
     print(fr)
     # Tabela de frequência (incluindo os valores nulos)
     fan = dados2021.SEXO.value_counts(dropna=False)
     print('Frequência absoluta considerando os nulos')
     print(fan)
     # Tabela de frequência relativa (incluindo os valores nulos)
     frn = dados2021.SEXO.value_counts(normalize = True, dropna=False)
     print('Frequência relativa considerando os nulos')
```

```
Frequência absoluta
Masculino     235779
Feminino      188606
Indefinido        45
Name: SEXO, dtype: int64
Frequência relativa
Masculino     0.555519
Feminino      0.444375
Indefinido    0.000106
Name: SEXO, dtype: float64
Frequência absoluta considerando os nulos
Masculino     235779
Feminino      188606
Indefinido        45
Name: SEXO, dtype: int64
Frequência relativa considerando os nulos
```

Fonte: (Autoria própria, 2023).

4.2 Discussão

A análise exploratória da base de dados de 2019 revelou que, antes do surgimento da pandemia de Covid-19, as doenças cardiovasculares já eram a principal causa de óbitos no Brasil. Isso contrasta com as desinformações e notícias falsas associadas às causas de óbito após a vacinação contra a Covid-19.

Os óbitos por Covid-19, cerca de 55.6% foram registrados entre os homens, enquanto 44,4% ocorreram entre as mulheres. Essa distribuição desigual pode ser influenciada por diversos fatores, incluindo características biológicas, comportamentais e socioeconômicas. Estudos têm indicado que os homens podem apresentar maior suscetibilidade a complicações graves relacionadas à Covid-19, o que pode contribuir para essa disparidade na mortalidade. Além disso, diferenças na exposição ao vírus, acesso a cuidados de saúde e adoção de medidas preventivas também podem desempenhar um papel nessa distribuição desigual.

Em relação à raça/cor, os dados mostram que a maior quantidade de óbitos por Covid-19 no Brasil em 2021 foi entre indivíduos da raça branca, representando aproximadamente 57% dos óbitos. Em segundo lugar, estão os indivíduos de raça parda, e em terceiro lugar, os de raça preta. Quanto ao local de ocorrência do óbito, a maioria dos óbitos por Covid-19 no Brasil em 2021 ocorreu em hospitais, seguido por outros estabelecimentos de saúde e, em terceiro lugar, em domicílio.

No que diz respeito à escolaridade, o maior percentual de óbitos por Covid-19 no Brasil em 2021 foi entre pessoas com ensino fundamental I. Em segundo lugar, estão aqueles com 8 a 11 anos de escolaridade, correspondente ao ensino médio (antigo 2º grau). Ao analisar a escolaridade em anos, verifica-se que a faixa mais comum de óbitos por Covid-19 no Brasil em 2021 foi entre aqueles com 8 a 11 anos de escolaridade, seguida pela faixa de 4 a 7 anos.

Na análise da distribuição de óbitos por sexo e idade em 2021, o gráfico *boxplot* da Figura 10 revela que a faixa etária entre 50 e 80 anos abriga a maioria dos óbitos para ambos os gêneros. Entretanto, é notável que o sexo feminino apresenta uma elevação mais significativa nas idades mais avançadas. A interpretação desse tipo de gráfico envolve a observação da caixa (*box*), que representa o intervalo interquartil (IQR), onde a maioria dos dados está concentrada. A linha no meio da caixa é a mediana, indicando a posição central dos dados, enquanto as "antenas" ou "bigodes" estendem-se até os valores extremos, excluindo *outliers*. Essa análise proporciona *insights* valiosos sobre a distribuição dos óbitos em diferentes faixas etárias e como ela difere entre os sexos.

Esses dados destacam a importância de considerar fatores demográficos e socioeconômicos na análise da distribuição de óbitos por Covid-19. As informações podem auxiliar na identificação de grupos mais vulneráveis e direcionar estratégias de prevenção e cuidado específicas para cada segmento da população.

5 Considerações Finais

Para as análises a ferramenta Python demonstrou ser uma ótima escolha de linguagem de programação para o desenvolvimento deste trabalho por seu potencial na execução de análises exploratórias de dados, englobando desde o processamento dos dados até a aplicação de estatística descritiva, criação de sumários estatísticos, elaboração de gráficos, utilização de técnicas de amostragem, aplicação da estatística inferencial e realização de cálculos de médias, medianas e correlações. A tecnologia, incluindo algoritmos preditivos e análises de dados, desempenham um papel crucial nesse cenário, fornecendo insights valiosos. Ao investir em abordagens proativas de cuidados de saúde, os gestores públicos ou de empresas privadas

podem não apenas melhorar a qualidade de vida da população, mas também reduzir custos associados ao tratamento de doenças em estágios avançados. Essa abordagem preventiva alinha-se não apenas aos interesses financeiros, mas também ao compromisso com a promoção da saúde e bem-estar da população atendida. O desenvolvimento deste trabalho foi realizado com Covid-19, mas esta análise é empregada para qualquer patologia informada como entrada de dados ou outras bases de dados, incentivando o uso da tecnologia nesse processo.

Referências

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis**. Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas e Agravos não Transmissíveis no Brasil 2021-2030**. Brasília: Ministério da Saúde, 2021. Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/doencas-cronicas-nao-transmissiveis-dcnt/09-plano-de-dant-2022_2030.pdf/@@download/file. Acesso em: 01 fev. 2025.

BRASIL. Ministério da Saúde**. Hipertensão afeta um a cada quatro adultos no Brasil**. Brasília: Ministério da Saúde, 2019. Disponível em: https://www.gov.br/saude/pt-br/assuntos/noticias/2019/abril/hipertensao-afeta-um-a-cada-quatro-adultos-no-brasil. Acesso em: 02 de dez. 2023.

BRASIL. Ministério da Saúde. **Mortalidade Geral - Estrutura. Diaad**. Disponível em: https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral+-+ Estrutura.pdf. Acesso em: 10 out. 2023.

FAWCETT, T.; PROVOST, F. **Data Science para Negócios**: O que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books Editora, 2018.

FELTRIN, F. **Tratamento de Dados com Python + Pandas.** São Paulo: Novatec Editora, 2021.

GASPAR, J. S. *et al*. **Introdução à Análise de Dados em Saúde com Python.** Belo Horizonte: Biblioteca J. Baeta Vianna/Universidade Federal de Minas Gerais, 2023. Disponível em: https://docs.bvsalud.org/biblioref/2023/06/1437637/introducao-a-analise-de-dados-em-saude-com-python-ciia-saude.pdf. Acesso em: 17 out. 2023.

HERNÁNDEZ SAMPIERI, R.; MENDOZA TORRES, C. P. **Metodología de la investigación**: Las rutas cuantitativa, cualitativa y mixta. Cidade de México: McGRAW-HILL Interamericana, 2018.

MACHADO, F. N. R. **Big Data**: o futuro dos dados e aplicações. São Paulo: Erica, 2018.

MATPLOTIB. **Matplotlib: Visualization with Python**. Matplotib. 2023. Disponível em: https://matplotlib.org/. Acesso em: 17 out. 2023.

McKINNEY, W. **Python para Análise de Dados**: Tratamento de Dados com Pandas, NumPy & Jupyter. São Paulo: Novatec Editora, 2023. Disponível em: https://wesmckinney.com/book/accessing-data. Acesso em: 29 out. 2023.

NETTO A, MACIEL F. **Python Para Data Science**: E Machine Learning Descomplicado. 1. ed. Rio de Janeiro: Alta Books Editora, 2021.

BRASIL. Ministério da Saúde. Sistema de Informação sobre Mortalidade - SIM. **OPEN DATASUS**. Disponível em:https://opendatasus.saude.gov.br/pt_BR/dataset/sim. Acesso em: 19 de fevereiro de 2023.

RIBEIRO NETO, J. A. **Big Data para Executivos e Profissionais de Mercado**. Jose Antonio Ribeiro Neto: Publicação independente, 2020.

SEABORN. Disponível em:https://seaborn.pydata.org/. Acesso em: 17 out. 2023.

TAN, P. N., STEINBACH, M., KUMAR, V., & KARPATNE, A. **Introduction to Data Mining: Global Edition**. 2. ed. Pearson Education, 2019.

WHO. World Health Organization. **Number of COVID-19 deaths reported to WHO**. Genebra: World Health Organization, 2019. Disponível em: https://data.who.int/dashboards/covid19/deaths?n=c/. Acesso em 13 Mai 2022.