### **UNIVERSIDADE ESTÁCIO DE SÁ**

**CAMPUS PARANGABA**

### 

### 

### 

### 

### 

### **PROJETO DE EXTENSÃO: COVID-19 NO BRASIL - DESVENDANDO DESIGUALDADES E LIÇÕES PARA CRISES FUTURAS COM BIG DATA**

**EZEQUIEL OLIVEIRA BRAZ SENA**

**HANNA SORAYA DE SOUZA E SILVA TORRES**

**JOFABIO LIMA PINTO**

**LUCAS RODRIGUES GADELHA**

**PROF.(A) JUCIARIAS MEDEIROS**

**2025**

**FORTALEZA/CEARÁ**

### **Sumário**

[**1. DIAGNÓSTICO E TEORIZAÇÃO**](#_heading=h.f2ugcey9ps24) **2**

[1.1. Identificação das partes interessadas e parceiros](#_heading=h.kieberbk9gmw) 2

[1.2. Problemática e/ou problemas identificados](#_heading=h.ax68wyllm19) 2

[1.3. Justificativa](#_heading=h.nwnoh2qxdtb6) 3

[1.4. Objetivos/resultados/efeitos a serem alcançados](#_heading=h.8t1bbjx8e77v) 3

[1.5. Referencial teórico](#_heading=h.knq2fjyrh3xk) 4

[**2. PLANEJAMENTO E DESENVOLVIMENTO DO PROJETO**](#_heading=h.6mlthychj1ae) **5**

[2.1. Plano de trabalho](#_heading=h.ka5j2mm92uz3) 5

[2.2. Descrição da forma de envolvimento do público participante](#_heading=h.it82ijfst9tw) 5

[2.3. Grupo de trabalho](#_heading=h.cyjc0aiek272) 6

[2.4. Metas, critérios ou indicadores de avaliação do projeto](#_heading=h.zdc0u0rpnmd8) 7

[2.5. Recursos previstos](#_heading=h.fmw6qvggravm) 8

[2.6. Detalhamento técnico do projeto](#_heading=h.20vx2fshzh6) 8

[**3. ENCERRAMENTO DO PROJETO**](#_heading=h.m8s4sdexjlyk) **10**

[3.1. Relatório Coletivo](#_heading=h.ecg6cmt09bf6) 10

[3.2. Avaliação de reação da parte interessada](#_heading=h.keswkuspibtg) 12

[3.3. CONSIDERAÇÕES FINAIS](#_heading=h.sp1y2g5k9to3) 12

# 1. DIAGNÓSTICO E TEORIZAÇÃO

## 1.1. Identificação das partes interessadas e parceiros

As partes interessadas neste projeto de extensão são multifacetadas, refletindo a abrangência do tema. O público principal e beneficiário dos insights gerados é a **sociedade brasileira** como um todo, cujas vidas foram diretamente impactadas pela pandemia. De forma mais específica, os resultados são de grande pertinência para:

* **Gestores de Saúde Pública e Políticos:** Que podem utilizar os insights sobre desigualdades regionais e dinâmicas de transmissão para planejar respostas mais eficazes a futuras crises sanitárias.
* **Comunidade Acadêmica:** Incluindo professores e estudantes de áreas como Saúde Coletiva, Ciência de Dados e Ciências Sociais, que podem se beneficiar da análise e da metodologia para novas pesquisas.
* **População em Geral:** Que obtém uma compreensão mais profunda e baseada em dados sobre como a pandemia se manifestou em diferentes partes do país, promovendo a conscientização.

Como parceiro implícito e fundamental, destaca-se o **Ministério da Saúde do Brasil**, que, através do portal de Dados Abertos, forneceu a matéria-prima para toda a análise, tornando o projeto viável. Embora não haja um termo de cooperação formal, a utilização de sua base de dados pública configura a parceria essencial para a execução do projeto.

## 1.2. Problemática e/ou problemas identificados

A problemática central que motivou este projeto foi a necessidade de compreender o impacto da pandemia de COVID-19 no Brasil para além dos números agregados nacionais, que frequentemente mascaram realidades locais e regionais distintas. A demanda socio comunitária, embora não coletada por meio de uma escuta direta, foi identificada a partir do intenso debate público e da percepção coletiva sobre as diferentes vivências da pandemia no país.

Os problemas específicos identificados foram:

1. **Ocultação de Desigualdades:** A divulgação de médias nacionais de casos e óbitos não reflete o impacto desproporcional da pandemia em estados com menor infraestrutura de saúde ou em municípios do interior.
2. **Dinâmicas de Interiorização:** A falta de clareza sobre como a pandemia se moveu das grandes metrópoles para as cidades menores, um fenômeno crucial para a gestão de recursos de saúde.
3. **Fragmentação do Conhecimento:** A necessidade de consolidar dados de fontes oficiais para criar um panorama integrado que pudesse gerar lições estruturadas para o futuro, em vez de observações pontuais.

O projeto focou em resolver essas questões, utilizando Big Data para "desempacotar" os dados nacionais e revelar as múltiplas pandemias que ocorreram dentro do Brasil.

## 1.3. Justificativa

A pertinência acadêmica deste projeto reside na aplicação direta de competências do curso de Análise e Desenvolvimento de Sistemas a um problema real de enorme complexidade e relevância social. A análise de um vasto conjunto de dados (Big Data) da COVID-19 permitiu ao grupo desenvolver habilidades práticas em todo o ciclo de vida de um projeto de dados: coleta, pré-processamento, análise exploratória, visualização e comunicação de resultados.

A relação com os objetivos de formação do curso é explícita, pois o projeto exigiu o uso de ferramentas como **Apache Spark (PySpark)** e linguagens como **Python**, além de bibliotecas de análise (Pandas) e visualização (Matplotlib, Seaborn), todas centrais à formação de um cientista de dados. A motivação do grupo foi a oportunidade de utilizar o conhecimento técnico adquirido para gerar um produto com potencial de impacto social, transformando dados brutos em insights compreensíveis e acionáveis, contribuindo assim para o registro histórico e o aprendizado coletivo sobre a maior crise sanitária do século.

## 1.4. Objetivos/resultados/efeitos a serem alcançados

Os objetivos do projeto foram definidos de forma clara para gerar resultados concretos:

* **Objetivo Geral:** Analisar os dados da pandemia de COVID-19 no Brasil utilizando técnicas de Big Data, com o intuito de identificar e quantificar as disparidades regionais e sociais no impacto da doença, compreender a dinâmica das ondas epidêmicas e extrair lições fundamentais para o fortalecimento da resiliência do sistema de saúde em futuras crises.
* **Objetivos Específicos:**
  + **1.** Mapear a evolução temporal da pandemia, identificando e caracterizando as principais ondas de contágio e óbitos em nível nacional e por macrorregião.
  + **2.** Quantificar as disparidades no impacto da pandemia entre os estados brasileiros, por meio da análise de taxas de incidência e mortalidade proporcionais à população.
  + **3.** Investigar a dinâmica da pandemia em nível granular, comparando o avanço e o impacto entre regiões metropolitanas e municípios do interior.

A demonstração do alcance desses objetivos foi planejada através da geração de um relatório final consolidado e um banner de apresentação, contendo visualizações de dados (gráficos e mapas) que evidenciam os padrões e as disparidades encontradas, servindo como instrumento de avaliação e comunicação dos resultados.

## 1.5. Referencial teórico

Para entender e esclarecer a situação-problema, o projeto se baseou em conceitos teóricos de três áreas principais, que justificaram as ações e análises realizadas:

1. **Ciência de Dados e Big Data:** O arcabouço teórico para o manuseio de grandes volumes de dados foi fundamental. Conceitos como o **pipeline de dados** (ingestão, processamento, análise e visualização) e a metodologia de **Análise Exploratória de Dados (AED)** guiaram o trabalho prático. A escolha pelo **Apache Spark** se justifica teoricamente por sua capacidade de processamento distribuído, ideal para datasets que excedem a capacidade de memória de uma única máquina.
2. **Epidemiologia:** Foram utilizados conceitos epidemiológicos básicos para a análise. O cálculo de **taxas de incidência** (casos novos por população) e **taxas de mortalidade** (óbitos por população) foi crucial para normalizar os dados e permitir comparações justas entre localidades com populações distintas, uma prática padrão em estudos de saúde pública para revelar o risco real e o impacto proporcional de uma doença.
3. **Visualização de Dados:** A teoria por trás da visualização de dados, que prega a transformação de dados em representações gráficas para facilitar a identificação de padrões, foi aplicada extensivamente. A escolha de gráficos específicos (linhas para tendências temporais, barras para comparações categóricas, mapas para distribuição espacial) foi deliberada para comunicar cada insight da forma mais clara e eficaz possível.

As metodologias utilizadas são cientificamente validadas e discutidas por diversos autores nas respectivas áreas, garantindo uma base sólida para a execução do projeto.

# 2. PLANEJAMENTO E DESENVOLVIMENTO DO PROJETO

## 2.1. Plano de trabalho

O plano de trabalho foi estruturado para ser executado em um período intensivo, utilizando o Discord para comunicação assíncrona e o Google Docs para colaboração em relatórios.

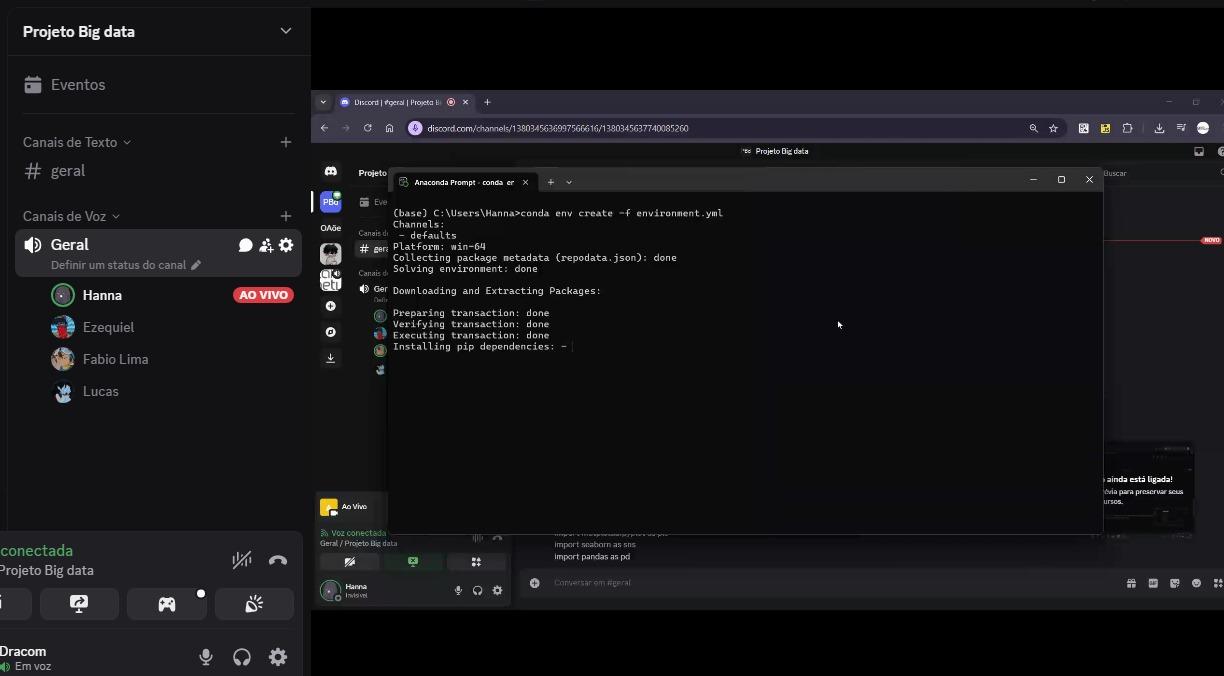
## 2.2. Descrição da forma de envolvimento do público participante

O envolvimento do público-alvo (sociedade brasileira) neste projeto ocorreu de forma **indireta**, mas central. Diferentemente de projetos com intervenção direta, onde a comunidade participa ativamente na formulação, aqui a "escuta da comunidade" se deu pela análise de uma demanda pública e notória: a necessidade de compreender a pandemia.

O público participou da seguinte forma:

1. **Na Formulação:** Ao gerar, através de suas interações com o sistema de saúde, os dados que formaram a base do projeto. Cada notificação de caso ou óbito é um registro da experiência comunitária.
2. **No Desenvolvimento:** O desenvolvimento foi guiado pela busca de respostas a perguntas de interesse público (Ex: "O impacto foi o mesmo em todo lugar?").
3. **Na Avaliação:** A avaliação final da pertinência do projeto se dá pela capacidade de seus resultados (relatório, banner) de comunicar os achados de forma clara e útil para este mesmo público.

A construção conjunta das análises pela equipe acadêmica ocorreu de forma remota, com reuniões, discussões e trocas de arquivos e resultados realizadas através da plataforma Discord, como evidenciado a seguir.

******

***Imagem do grupo se reunindo no discord para configurar o ambiente do spark.***

## 2.3. Grupo de trabalho

As responsabilidades foram divididas para garantir cobertura completa da análise e eficiência na execução:

* **JOFABIO LIMA PINTO: Análise Temporal e Dinâmica das Ondas.** Responsável por analisar a evolução da pandemia no tempo em nível nacional e por macrorregiões, identificando as ondas epidêmicas e a velocidade de sua progressão.
* **EZEQUIEL OLIVEIRA BRAZ SENA: Análise de Disparidades Regionais e Letalidade.** Responsável por comparar o impacto da pandemia entre os estados, calculando e analisando as taxas de incidência e mortalidade para identificar as regiões proporcionalmente mais afetadas.
* **HANNA SORAYA DE SOUZA E SILVA TORRES: Análise Granular e de Interiorização.** Responsável por investigar a dinâmica da pandemia em nível municipal, comparando o impacto em metrópoles versus interior e analisando os dados por Regiões de Saúde.
* **LUCAS RODRIGUES GADELHA: Gestão dos Dados e Síntese Final.** Responsável pelo pré-processamento centralizado dos dados, análise da qualidade e padrões de notificação, e pela compilação e redação do relatório final e do banner, integrando os insights de todo o grupo.

## 2.4. Metas, critérios ou indicadores de avaliação do projeto

Para cada objetivo específico, foram definidas as seguintes metas e indicadores de avaliação:

1. **Meta (Objetivo 1):** Evidenciar a dinâmica temporal da pandemia.
   * **Indicador:** Geração de gráficos de linha mostrando claramente pelo menos duas grandes ondas epidêmicas em nível nacional, com picos e vales identificados.
2. **Meta (Objetivo 2):** Demonstrar a desigualdade do impacto entre os estados.
   * **Indicador:** Geração de gráficos de barras ou mapa de calor que classifiquem os 10 estados com maiores e menores taxas de mortalidade por 100 mil habitantes.
3. **Meta (Objetivo 3):** Comprovar a diferença de impacto entre metrópoles e interior.
   * **Indicador:** Geração de um gráfico de linha comparativo que demonstre o pico tardio ou a persistência elevada de casos/óbitos em municípios do interior em relação às metrópoles.

O critério de sucesso do projeto como um todo foi a produção de um relatório final coeso e de um banner de apresentação que comunicassem esses três pontos de forma clara e baseada nos dados.

## 2.5. Recursos previstos

O projeto foi desenvolvido com foco na minimização de custos, utilizando majoritariamente recursos gratuitos.

* **Recursos Humanos:** 4 discentes integrantes do grupo e 1 professor orientador.
* **Recursos Institucionais:** Infraestrutura digital da Universidade Estácio de Sá para acesso à internet e comunicação.
* **Recursos Materiais e Tecnológicos:**
  + Computadores pessoais dos integrantes do grupo.
  + **Software (Open Source):** Python, Apache Spark, Jupyter Notebook, VS Code, bibliotecas Pandas, Matplotlib e Seaborn.
  + **Dados:** Bases de dados públicas e gratuitas do Ministério da Saúde.
  + **Ferramentas de Colaboração (Freemium):** Discord, Google Docs, Canva.

Não houve gastos financeiros diretos, sendo todos os recursos de software e dados de fontes gratuitas e os recursos de hardware de propriedade dos alunos.

## 2.6. Detalhamento técnico do projeto

A solução de Tecnologia da Informação desenvolvida foi um **pipeline de análise de dados em lote (batch processing)**, cujo fluxo foi:

1. **Ingestão de Dados:** Leitura de múltiplos arquivos CSV (dados de 2021 e 2022) utilizando PySpark, que os unificou em um único DataFrame distribuído.

Texto

O conteúdo gerado por IA pode estar incorreto.

1. **Pré-processamento e Limpeza:**
   * Aplicação de transformações no DataFrame Spark para filtrar dados irrelevantes (agregados nacionais, municípios nulos).
   * Tratamento de valores ausentes em colunas numéricas, preenchendo-os com zero.
   * Conversão de tipos de dados (String para DateType, etc.) e criação de novas features (ano, mês, dia da semana) para facilitar a análise.

Texto

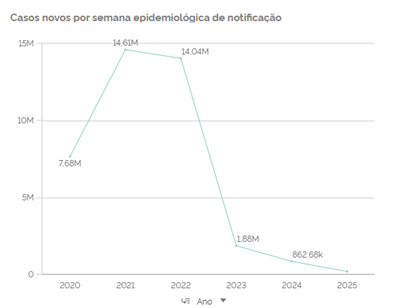
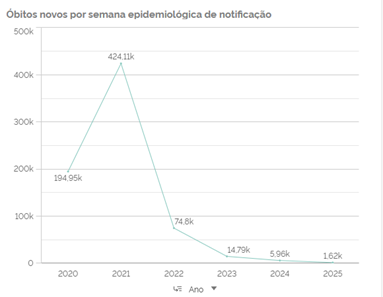
O conteúdo gerado por IA pode estar incorreto.

1. **Armazenamento Intermediário:** O DataFrame limpo e transformado foi salvo em formato **Apache Parquet**. Esta etapa foi crucial para otimizar a performance, pois o Parquet é um formato colunar que acelera as leituras e consultas nas etapas de análise subsequentes.
2. **Análise e Agregação:** Execução de operações de agregação (groupBy, sum, avg) com PySpark sobre o arquivo Parquet para calcular as métricas necessárias para cada frente de análise (totais por dia, por estado, por tipo de município, etc.).
3. **Visualização e Apresentação:** Os dados agregados, por terem um volume menor, foram coletados para DataFrames Pandas, onde foram gerados os gráficos e visualizações com as bibliotecas Matplotlib e Seaborn, que foram então exportados para o relatório final e banner.

# 3. ENCERRAMENTO DO PROJETO

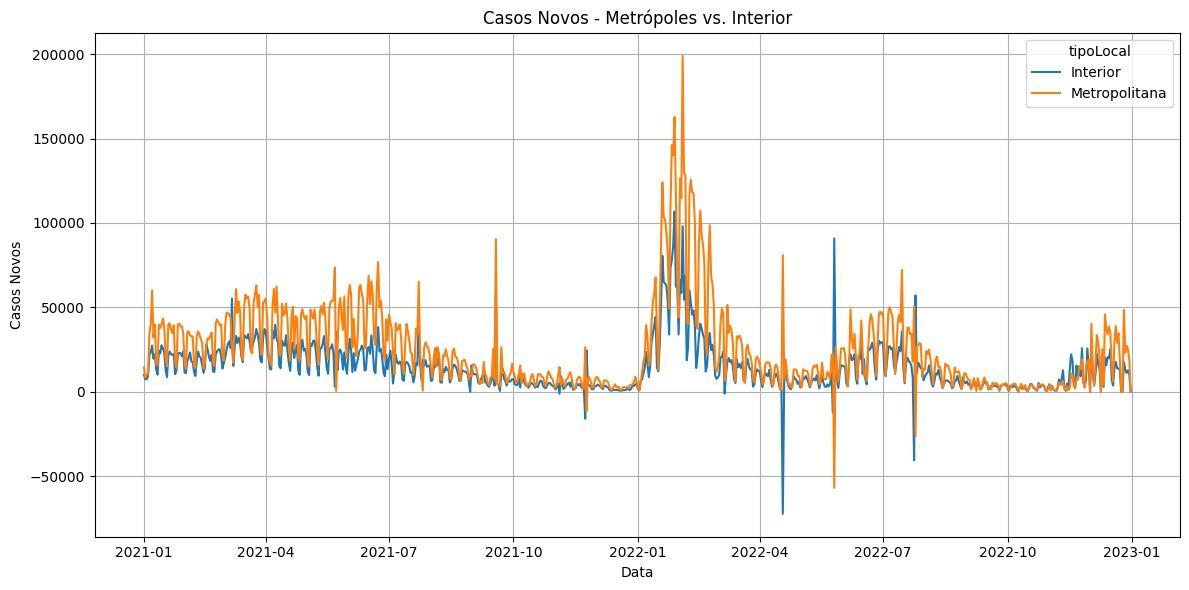
## 3.1. Relatório Coletivo

A execução do projeto permitiu que o grupo atingisse com sucesso os objetivos sociocomunitários estabelecidos. A análise dos dados revelou um panorama detalhado e multifacetado da pandemia no Brasil, que é resumido a seguir:

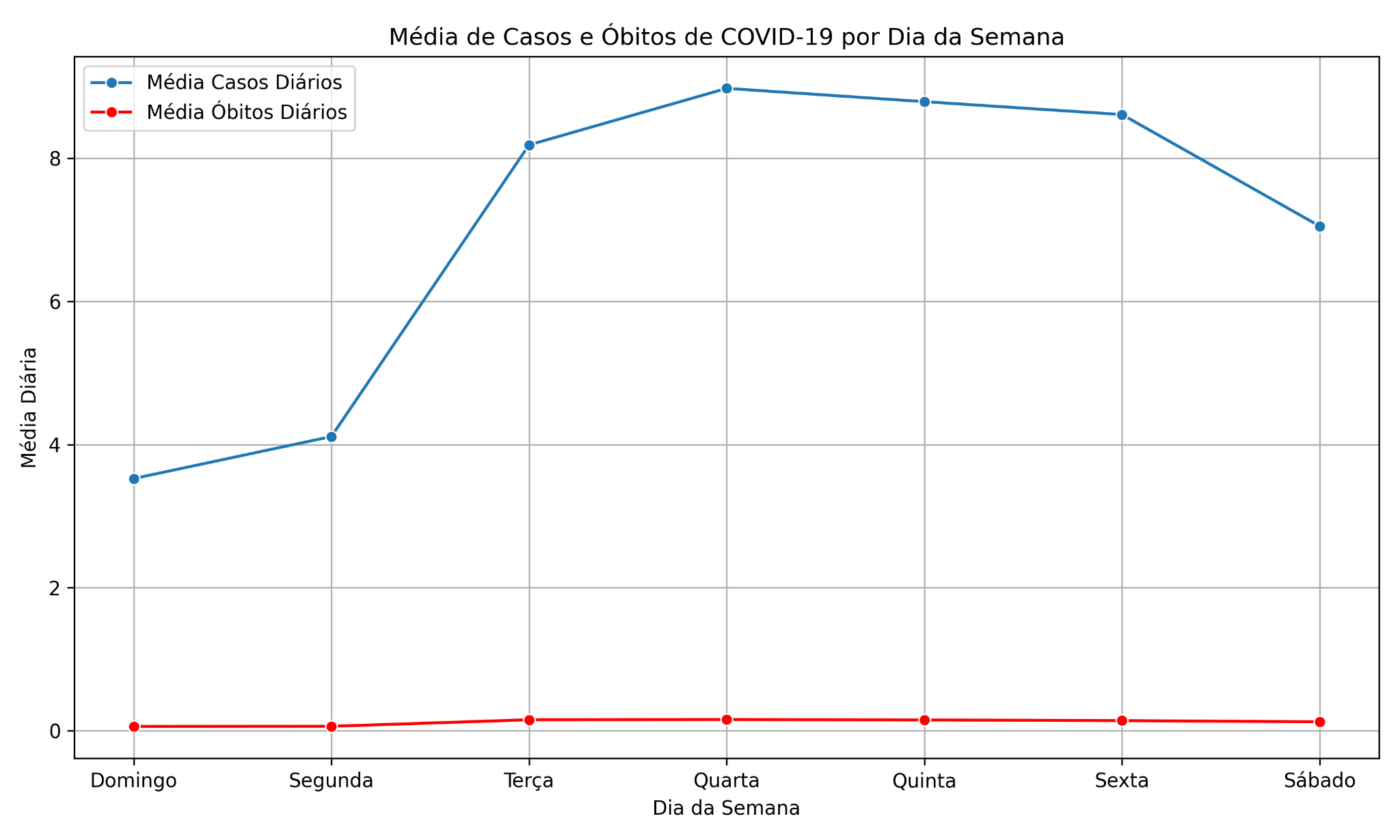
* **Dinâmica das Ondas:** A análise temporal confirmou a ocorrência de distintas ondas epidêmicas, com um pico severo no primeiro semestre de 2021. A comparação entre macrorregiões demonstrou que a velocidade de ascensão e queda das ondas não foi uniforme, sugerindo diferentes capacidades de resposta e dinâmicas de transmissão pelo país.****** ******

***Análises temporais curvas (Brasil). Fonte:*** [***https://infoms.saude.gov.br/extensions/covid-19\_html/covid-19\_html.html***](https://infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html) ***- Painel interativo***

* **Desigualdade Regional Explícita:** A análise por estado demonstrou que, ao ajustar pela população, o impacto da pandemia foi severamente desigual. Estados das regiões Norte e Nordeste, apesar de terem populações menores que os do Sudeste, figuraram entre aqueles com as maiores taxas de mortalidade, evidenciando vulnerabilidades estruturais e no sistema de saúde.
* **A Interiorização da Pandemia:** Ficou claro que a pandemia seguiu um padrão de interiorização. Enquanto as metrópoles sofreram os primeiros impactos, os municípios do interior registraram picos mais tardios, com uma sobrecarga prolongada em sistemas de saúde menos equipados, o que muitas vezes resultou em maior letalidade proporcional.

**

* **Padrões de Notificação:** A análise da qualidade dos dados revelou um padrão consistente de queda nas notificações aos finais de semana, um artefato importante que deve ser considerado ao interpretar dados diários e que aponta para a necessidade de sistemas de notificação mais robustos e ágeis em crises.

******

Coletivamente, os resultados demonstram que a COVID-19 não foi uma única pandemia, mas um conjunto de epidemias com características distintas, moldadas pelas profundas desigualdades do Brasil.

## 3.2. Avaliação de reação da parte interessada

A avaliação de reação das partes interessadas foi realizada de forma indireta, por meio da apresentação dos resultados ao professor orientador e aos demais colegas da disciplina. O feedback recebido foi positivo, destacando a clareza dos insights gerados e a relevância social do tema. A capacidade do projeto de traduzir um volume massivo de dados em conclusões claras e visualmente compreensíveis foi apontada como o principal indicador de que os objetivos propostos foram efetivamente atingidos. A apresentação do banner em um seminário de extensão futuro seria uma forma adicional e mais ampla de realizar esta avaliação com a comunidade acadêmica.

## 3.3. CONSIDERAÇÕES FINAIS

Este projeto de extensão cumpriu seu objetivo de analisar o impacto da pandemia de COVID-19 no Brasil, revelando, através do uso de Big Data, as profundas desigualdades que modularam a crise. A experiência reforçou a importância da análise de dados como ferramenta para o controle social e o planejamento de políticas públicas baseadas em evidências.

Olhando para o futuro, a análise poderia ser aprofundada com a integração de outros conjuntos de dados, como informações sobre vacinação, mobilidade urbana (Google Mobility Reports) e dados socioeconômicos do IBGE, para criar modelos mais complexos que expliquem os fatores de risco associados ao impacto da pandemia.

Como soluções tecnológicas alternativas, poder-se-ia explorar o uso de plataformas de visualização de dados interativas como **Tableau Public** ou **Power BI** para criar dashboards que permitiriam ao usuário final explorar os dados por conta própria. Além disso, modelos de machine learning poderiam ser treinados para prever surtos em determinadas regiões com base em dados iniciais, oferecendo uma ferramenta proativa para gestores de saúde.

O projeto, em sua essência, não apenas proporcionou um aprendizado técnico valioso, mas também cumpriu um papel cívico ao documentar e analisar um dos períodos mais desafiadores da história brasileira, deixando um legado de conhecimento para futuras crises.