**UNIVERSIDADE ESTÁCIO DE SÁ  
CAMPUS PARANGABA**

**CIÊNCIA DA COMPUTAÇÃO**

**LUCAS RODRIGUES GADELHA**

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

**FORTALEZA – CEARÁ  
2025**

**LUCAS RODRIGUES GADELHA**

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

Relatório acadêmico apresentado à disciplina “Tópicos de Big Data em Python” do curso de Análise e Desenvolvimento de Sistemas da Universidade Estácio de Sá, como requisito parcial para avaliação.

**Orientador(a): Prof.(a) Juciarias Medeiros**

**FORTALEZA – CEARÁ  
2025**

**SUMÁRIO**

[**1. CONTEXTUALIZAÇÃO**](#_heading=h.ckg3qimgzdue) **3**

[**2. OBJETIVOS**](#_heading=h.2dacm75qapwj) **3**

[**3. METODOLOGIA**](#_heading=h.3bzbp03az0p) **3**

[**4. RESULTADOS E DISCUSSÃO**](#_heading=h.xrd55s79a37s) **4**

[**5. REFLEXÃO APROFUNDADA**](#_heading=h.4vrwdu2zddia) **5**

# 1. CONTEXTUALIZAÇÃO

A experiência vivida foi a participação no projeto de extensão "COVID-19 no Brasil: Desvendando Desigualdades e Lições para Crises Futuras com Big Data", realizado no âmbito da disciplina “Tópicos de Big Data em Python”. Neste projeto, atuei no papel de "Pessoa 4", uma função central que envolvia a gestão de dados e a síntese final dos resultados do grupo. Minha participação consistiu em receber os dados brutos, realizar todo o pré-processamento para criar uma base de dados limpa e otimizada para a equipe, analisar a qualidade e a consistência dessa base, e, por fim, integrar as análises individuais dos meus colegas em um relatório final coeso e em um banner de apresentação.

# 2. OBJETIVOS

Os objetivos específicos da minha participação e experiência no projeto foram:

* **Preparar um Dataset Confiável:** Garantir que todo o grupo trabalhasse sobre uma base de dados única, limpa e consistente, tratando valores nulos, corrigindo tipos de dados e otimizando o acesso através da conversão para o formato Parquet.
* **Analisar a Qualidade dos Dados:** Investigar a fundo a base de dados bruta para identificar padrões, vieses e anomalias que pudessem impactar as análises, com foco principal na consistência das notificações ao longo do tempo.
* **Sintetizar e Integrar os Resultados:** Compilar os insights gerados por todas as frentes de análise (temporal, regional e granular) em uma narrativa lógica e unificada, materializada no relatório final e no banner de apresentação do projeto.

# 3. METODOLOGIA

A experiência foi vivenciada de forma remota, tendo como local de trabalho principal o ambiente de desenvolvimento em nossas próprias máquinas. Os sujeitos envolvidos foram os quatro discentes do grupo e o professor orientador, que nos guiava remotamente. A comunicação e a colaboração da equipe ocorreram por meio de canais de texto e voz na plataforma Discord, durante um período de uma semana, entre 03 e 07 de junho de 2025.

O detalhamento das minhas etapas de trabalho foi o seguinte:

1. **Ingestão e Consolidação dos Dados:** Iniciei carregando múltiplos arquivos CSV dos anos de 2021 e 2022 com PySpark.
2. **Limpeza e Pré-processamento:** Filtrei registros desnecessários, tratei valores ausentes em colunas numéricas (casos e óbitos) e converti os tipos de dados para os formatos corretos (ex: data).
3. **Otimização para a Equipe:** Salvei o DataFrame limpo no formato Apache Parquet, o que reduziu o tempo de leitura e garantiu que todos os membros do grupo tivessem acesso rápido à mesma base de dados.
4. **Análise de Qualidade:** Executei rotinas para verificar a presença de duplicatas e para analisar a consistência dos dados acumulados. Agreguei os totais de casos e óbitos por dia da semana para investigar os padrões de notificação.
5. **Síntese dos Achados:** Reuni os relatórios e gráficos individuais de cada colega, organizei os insights e redigi o relatório final e o banner.

# 4. RESULTADOS E DISCUSSÃO

A minha expectativa inicial era a de que os dados, por serem de uma fonte oficial, seriam majoritariamente consistentes. O vivido, contudo, revelou a complexidade de dados gerados em um cenário de crise. O que observei foi que, embora os dados fossem robustos, eles continham vieses intrínsecos ao processo de notificação.

O resultado mais significativo da minha análise foi a descoberta do "efeito do fim de semana": uma queda sistemática nas notificações de casos e óbitos aos sábados e domingos, seguida por um pico de compensação nas terças-feiras. A experiência de descobrir e quantificar esse padrão foi reveladora. Me senti como um detetive, pois este achado mudou a forma como o grupo inteiro interpretou os dados de tendências diárias.

As dificuldades incluíram o gerenciamento do volume inicial dos arquivos CSV e a coordenação para que todos usassem a versão correta do dataset tratado. A principal facilidade foi o uso do formato Parquet, que, após o processamento inicial, tornou as análises subsequentes muito mais rápidas para todos. A maior descoberta/aprendizagem foi, sem dúvida, a importância crítica da etapa de engenharia e análise de qualidade de dados como alicerce para qualquer projeto de ciência de dados.

Como **recomendações** para projetos futuros, sugiro:

* Aprofundar a análise de qualidade com métodos estatísticos para detecção automática de anomalias.
* Considerar o desenvolvimento de um dashboard de monitoramento de dados (com ferramentas como Streamlit ou Dash) para que a equipe possa acompanhar a qualidade dos dados de forma interativa.

# 5. REFLEXÃO APROFUNDADA

A relação entre a experiência vivida e a teoria estudada foi direta e tangível. Na disciplina, discutimos o conceito de "pipeline de dados" de forma teórica; no projeto, eu fui o responsável por construir e executar as etapas iniciais desse pipeline (ingestão, limpeza, transformação). A teoria sobre Big Data aponta para a necessidade de ferramentas como o Apache Spark para lidar com grandes volumes de dados, e eu vivenciei essa necessidade na prática ao processar os arquivos CSV que eram lentos de manipular com ferramentas tradicionais.

Além disso, a teoria da análise de dados enfatiza a importância de não aceitar os dados como uma verdade absoluta, mas de investigá-los criticamente. Minha análise do "efeito do fim de semana" foi a personificação dessa teoria. Em vez de apenas usar os dados, eu os questionei, resultando em um insight que enriqueceu e trouxe mais precisão a todas as outras análises do grupo. A experiência, portanto, serviu para solidificar os conceitos teóricos, transformando-os de abstrações em ferramentas práticas e indispensáveis.