UNIVERSIDADE FEDERAL DO CEARÁ

ENGENHARIA AMBIENTAL E SANITÁRIA

DOCENTE RESPONSÁVEL: RENAN VIEIRA GOMES

LUIZ GUSTTAVO MACEDO MAGALHÃES

RELATÓRIO TRABALHO 2 CIÊNCIA DE DADOS:

ANÁLISE EXPLORATÓRIA DE DADOS (AED) E VISUALIZAÇÕES

CRATEÚS

2025

1. Objetivo

Este trabalho teve como objetivo principal realizar uma Análise Exploratória de Dados (AED) nos três conjuntos de dados selecionados no T1 (Iris, Heart Disease e COVID-19). O foco foi aplicar técnicas de limpeza e transformação de dados, calcular estatísticas descritivas e, principalmente, gerar visualizações (gráficos) para formular e responder perguntas sobre os dados, contando a história por trás deles (Storytelling).

2. Dados Utilizados

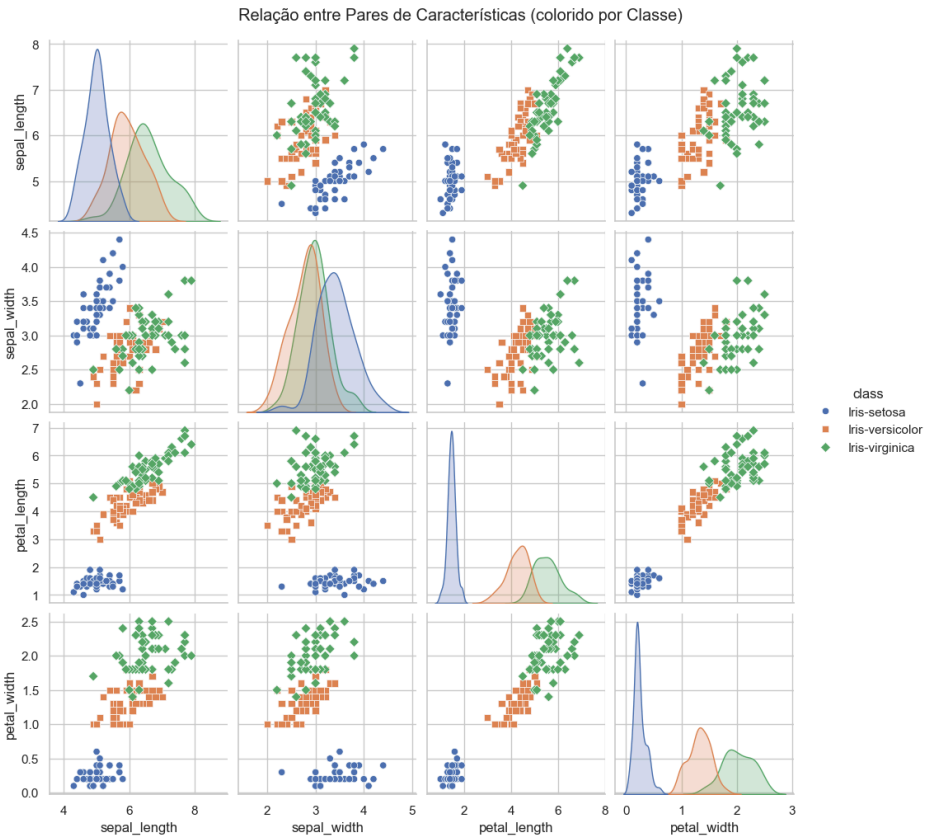
Foram utilizados os mesmos três datasets do T1, com as seguintes características de análise:

- **Iris:** Dataset clássico e limpo, usado para praticar a análise estatística e visualização (Histogramas, Box Plots, Pair Plots).
- **Heart Disease:** Dataset que exigiu **limpeza de dados**, tratando valores faltantes (?) que foram convertidos para NaN e depois removidos (dropna()). Também exigiu **transformação de tipos**, convertendo colunas numéricas (ex: sex, cp) para o tipo category e simplificando a coluna alvo num para um formato binário (0 ou 1).
- **COVID-19 (Brasil.IO):** Dataset grande, carregado via URL, que exigiu **transformação de dados** (conversão da coluna date para datetime no carregamento) e **filtragem** (selecionando apenas place_type == 'state' para uma análise macro). A análise focou em séries temporais (gráficos de linha) e agregações por estado.

3. Principais Achados e Visualizações

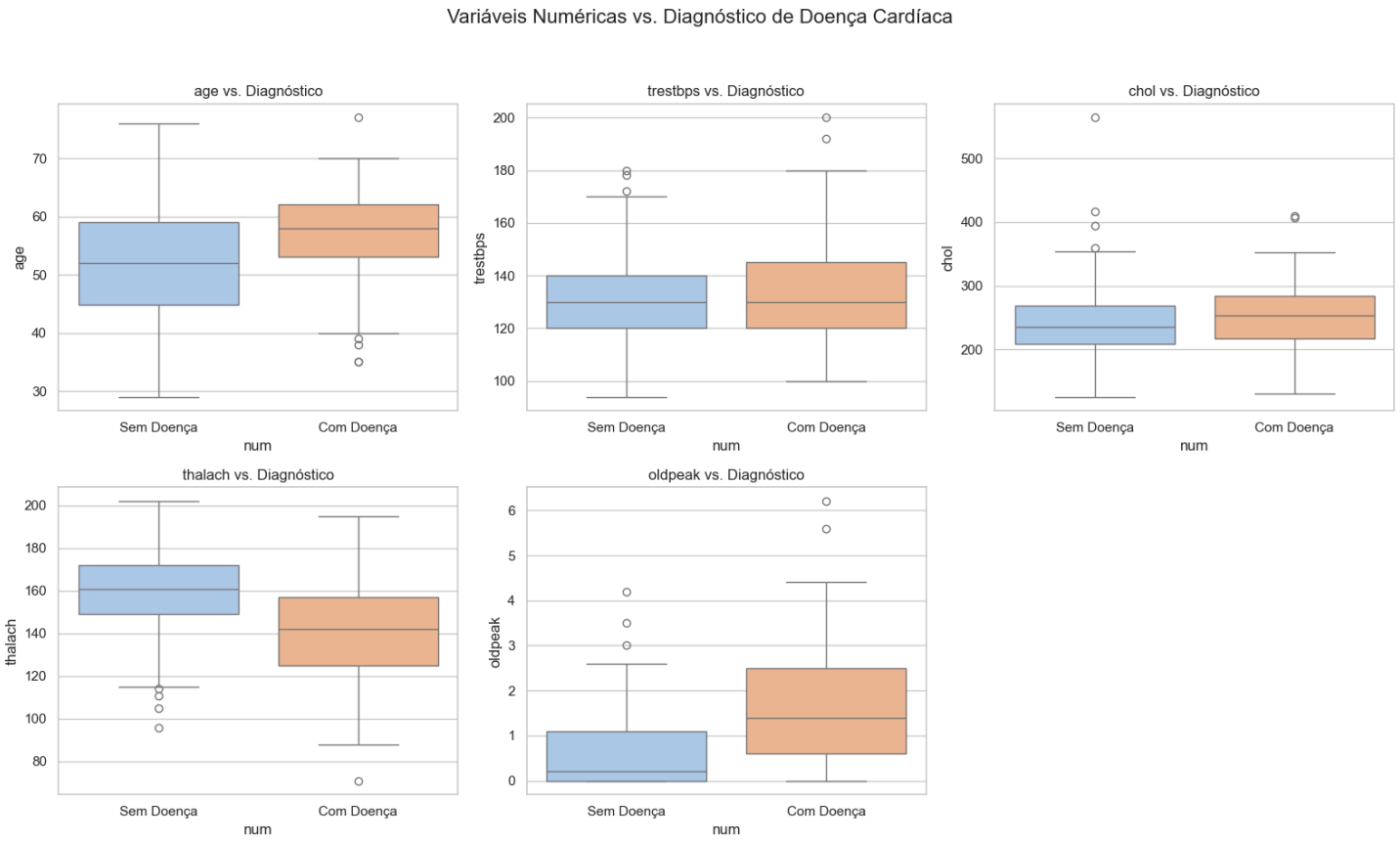
Abaixo estão os 4 principais "achados" (insights) descobertos durante a análise dos notebooks.

Achado 1: Iris - Separação Clara das Espécies pela Pétala



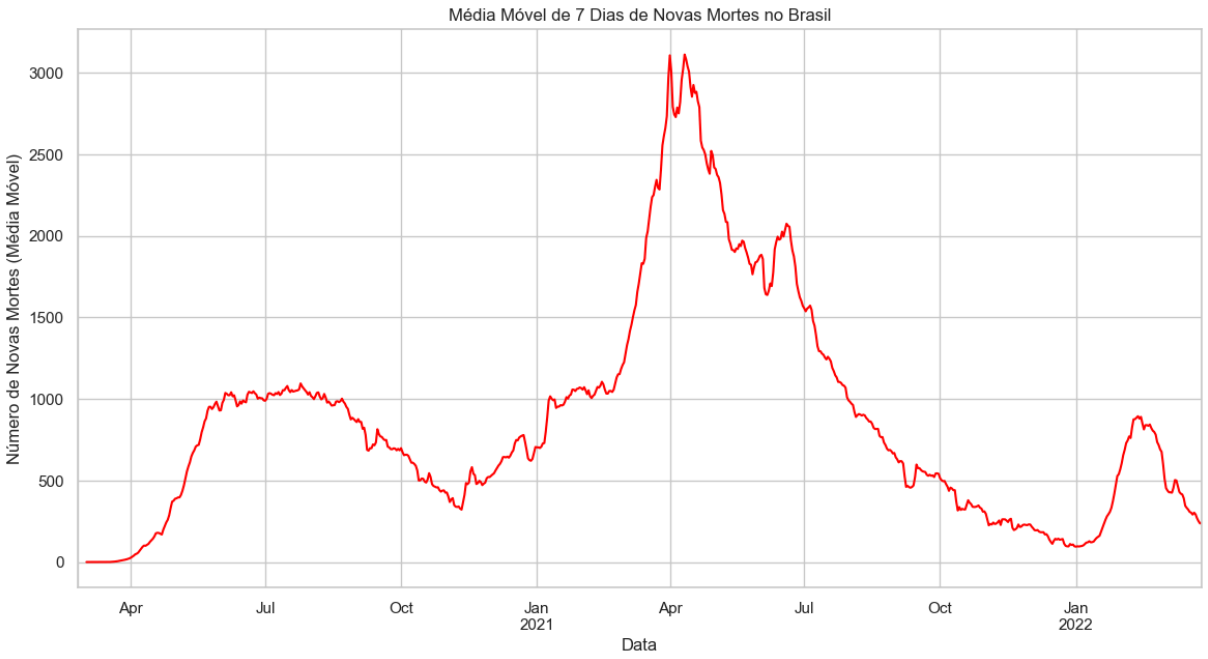
Pergunta: É possível separar a espécie 'Iris-setosa' das outras usando apenas as medidas da pétala?
Resposta: Sim. A visualização (Pair Plot) mostra que a 'Iris-setosa' forma um grupo completamente isolado com base no comprimento (petal_length) e largura (petal_width) da pétala, que são visivelmente menores que os das outras duas espécies.

Achado 2: Heart Disease - Frequência Cardíaca vs. Diagnóstico



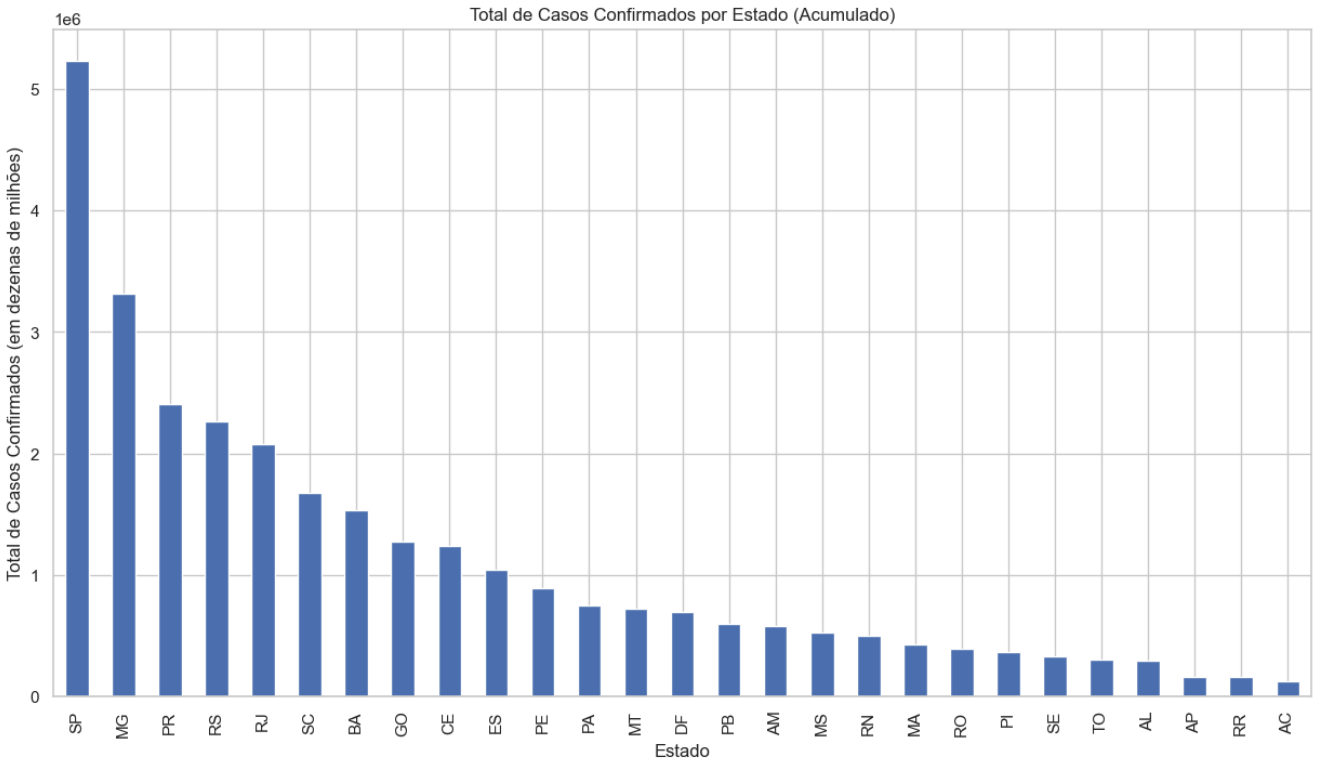
Pergunta: Pacientes com doença cardíaca (num=1) tendem a ter uma frequência cardíaca máxima (thalach) menor?
Resposta: Sim. O Box Plot demonstra claramente que a mediana da frequência cardíaca máxima em pacientes diagnosticados com a doença é significativamente mais baixa do que a mediana de pacientes saudáveis (num=0).

Achado 3: COVID-19 - Pico de Mortalidade da Pandemia



Pergunta: Observando a curva nacional, qual foi o período aproximado do pico de novas mortes por COVID-19 no Brasil?
Resposta: O gráfico de média móvel de 7 dias mostra que o pico mais agudo e letal da pandemia ocorreu no primeiro semestre de 2021, atingindo seu máximo por volta de abril de 2021.

Achado 4: COVID-19 - Distribuição de Casos por Estado



Pergunta: Qual estado brasileiro registrou o maior número acumulado de casos de COVID-19?
Resposta: O gráfico de barras mostra que o estado de São Paulo (SP) foi o que registrou o maior número de casos acumulados, com uma margem considerável em relação ao segundo colocado, Minas Gerais (MG).

4. Desafios e Aprendizados

O principal desafio foi lidar com as diferentes características de cada dataset. No T1, a dificuldade foi a organização dos arquivos e a configuração do ambiente, incluindo a decisão de como lidar com o arquivo de +100MB da COVID-19.

No T2, os principais aprendizados foram:

- **Limpeza de Dados:** A importância de tratar dados faltantes, como os ? no dataset *Heart Disease*, usando dropna() ou outras técnicas.
- **Transformação de Tipos:** A necessidade de converter colunas para seus tipos corretos (ex: parse_dates=['date'] no COVID e .astype('category') no Heart Disease) para que as análises e gráficos funcionem corretamente.
- **Visualização como Ferramenta:** Como as bibliotecas Matplotlib e Seaborn não servem apenas para "fazer gráficos bonitos", mas são ferramentas essenciais para *responder perguntas* e contar a história dos dados (Storytelling).

5. Limitações e Recomendações

Uma limitação notável é a idade dos datasets Iris (1936) e Heart Disease (1988); embora sejam clássicos para estudo, os dados podem não refletir cenários atuais. O dataset da COVID-19, por ser de origem pública, pode conter subnotificações.

Os dados limpos e transformados do Heart Disease e Iris estão agora prontos para a aplicação de modelos de Machine Learning (como Regressão Logística ou Árvores de Decisão) para prever o diagnóstico ou a espécie da flor.