

# **Relatório Técnico: Análise Exploratória de Dados Socioambientais na Amazônia**

**Curso:** IA voltado para Sustentabilidade para COP30

**Aluno:** Eric Pimentel

**Assunto:** Relatório Técnico da Tarefa Individual III

**Data:** 29 de Junho de 2025

## **1. Introdução e Definição do Problema**

O presente relatório detalha o processo de Análise Exploratória de Dados (EDA) realizado em um contexto socioambiental na região amazônica. O objetivo central do projeto foi transformar dados brutos, provenientes de duas fontes distintas (climática e socioeconômica), em insights acionáveis que pudessem servir de base para futuras soluções de sustentabilidade, como o desenvolvimento de modelos preditivos e a formulação de políticas públicas locais.

A análise foi guiada por um problema central, derivado de um cenário de crescente instabilidade hídrica e seus impactos percebidos pela população local:

- **Problema Central:** Quantificar e validar a correlação entre as variações climáticas e hídricas (escassez, excesso e qualidade da água) e seus impactos diretos na produtividade agrícola e na saúde (segurança alimentar e incidência de doenças) das comunidades amazônicas.

A partir disso, foram formuladas as seguintes questões de análise (hipóteses) para nortear a investigação:

1. Qual a natureza da correlação entre a precipitação pluviométrica e o volume da produção agrícola?
2. Existe uma correlação estatisticamente significativa entre a falta de acesso à água potável e a incidência de doenças de veiculação hídrica?
3. Qual variável apresenta maior impacto sobre o indicador de segurança alimentar: a instabilidade das chuvas ou a dificuldade de acesso à água

potável?

## 2. Metodologia: Estratégia de Preparação e Análise dos Dados

A metodologia foi estruturada para garantir a robustez e a confiabilidade dos resultados, seguindo as melhores práticas de ciência de dados. A abordagem combinou o uso de ferramentas de programação (Python e suas bibliotecas) com a aplicação de inteligência artificial generativa como assistente de codificação e análise ("vibe coding").

### 2.1. Estratégia de Limpeza e Pré-processamento de Dados

A qualidade da análise depende diretamente da qualidade dos dados. Portanto, uma etapa rigorosa de preparação foi executada, conforme detalhado abaixo:

- **Fusão de Dados:** As bases `base_climatica.csv` e `base_socioeconomica.csv` foram unificadas em um único DataFrame utilizando a coluna `data` como chave, após a conversão de ambas para o formato `datetime`.
- **Remoção de Duplicatas:** Foram identificados e removidos 20 registros duplicados, garantindo que cada observação fosse única.
- **Padronização de Variáveis Categóricas:** Os valores nas colunas `variacao_climatica` e `acesso_agua_potavel` foram normalizados para o formato minúsculo. Respostas como "nao" foram padronizadas para "não", eliminando ambiguidades.
- **Tratamento de Dados Ausentes (Missing Values):** A estratégia adotada foi a imputação de dados. Para variáveis numéricas, optou-se pela **mediana** em vez da média, por ser uma medida de tendência central mais robusta a outliers. Para variáveis categóricas, utilizou-se a **moda** (valor mais frequente).
- **Tratamento de Outliers:** Foi identificado um outlier significativo na variável `chuvas_reais_mm` (registros acima de 700mm, fisicamente improváveis para uma medição diária). Com base no contexto amazônico, onde eventos extremos podem chegar a 200-250mm, foi definida uma estratégia de **capping (limitação)**, ajustando todos os valores acima de 250mm para este teto. Essa

abordagem preserva a informação de que foi um evento extremo, sem distorcer a escala da análise.

## 2.2. Ferramentas Utilizadas

- **Linguagem:** Python 3.
- **Bibliotecas Principais:** Pandas (para manipulação de dados), Matplotlib e Seaborn (para visualização de dados).
- **Ambiente de Desenvolvimento:** Jupyter Notebook / Google Colab.

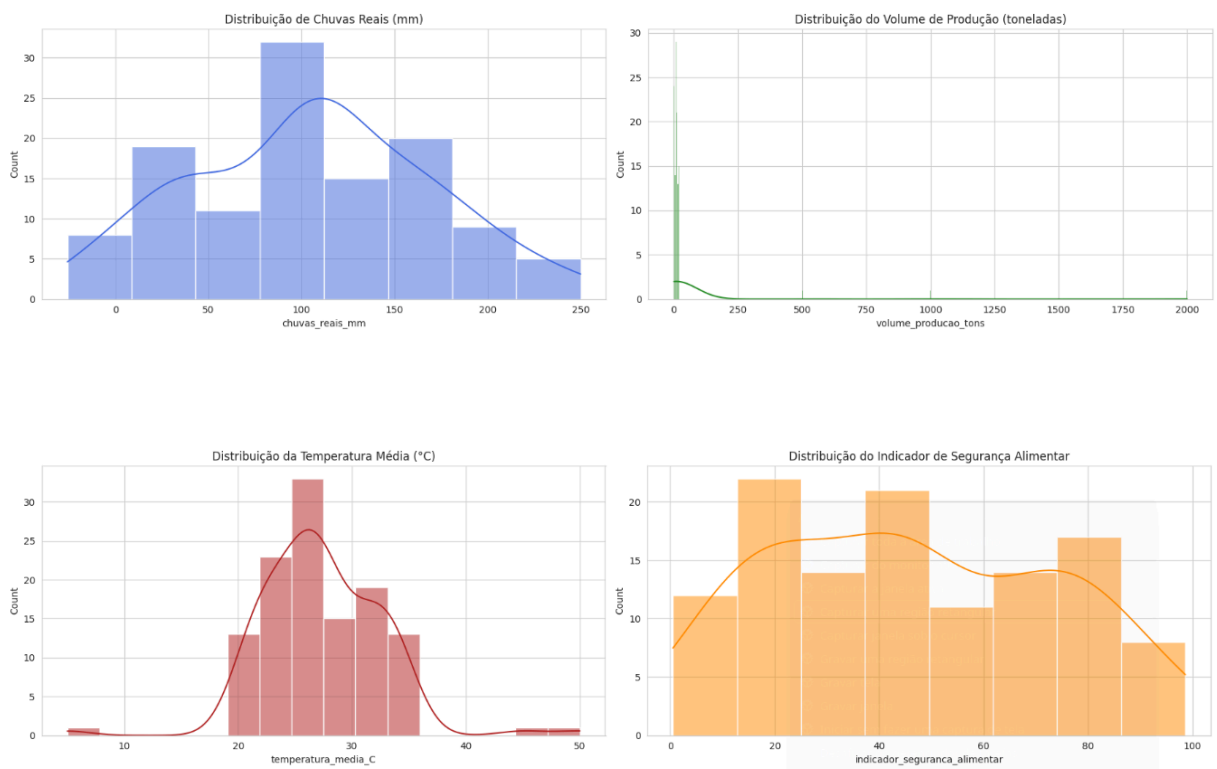
## 3. Análise Exploratória de Dados (EDA): Resultados e Interpretações

Nesta fase, buscou-se visualizar os dados para identificar padrões, testar as hipóteses e descobrir novas relações.

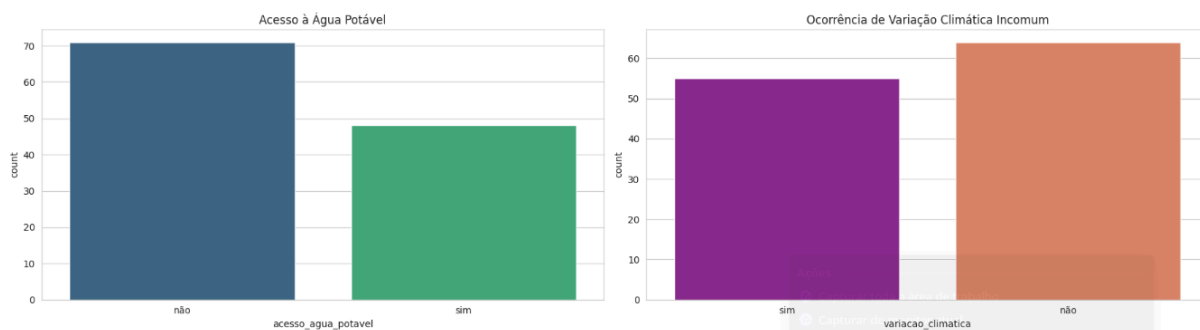
### 3.1. Análise Univariada: Distribuição das Variáveis-Chave

- Gráficos: Distribuição de Chuvas, Produção, Temperatura e Segurança Alimentar

Distribuição das Variáveis-Chave



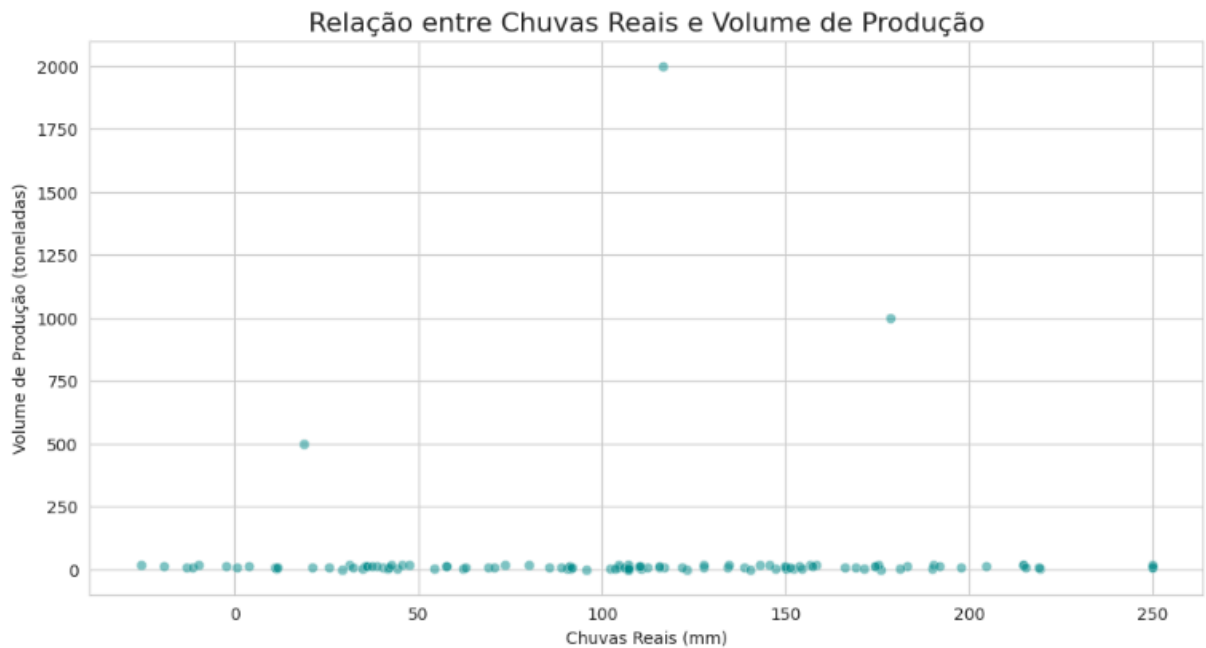
- **Chuvas Reais:** Apresenta uma distribuição concentrada entre 50mm e 150mm, o que é esperado para a região.
  - **Volume de Produção:** A distribuição é fortemente assimétrica, com a grande maioria dos registros concentrada em valores baixos e alguns poucos pontos representando picos de produção muito elevados.
  - **Temperatura Média:** Segue uma distribuição aproximadamente normal, centrada em torno de 25-30°C.
  - **Segurança Alimentar:** A distribuição é relativamente uniforme, sugerindo uma grande variação no nível de segurança alimentar entre as observações.
- Gráfico: Contagem de Acesso à Água e Variação Climática



Interpretação: A análise de contagem revela um dado social crítico: a maioria dos registros (~70) indica a falta de acesso à água potável, em comparação com um número menor (~50) com acesso.

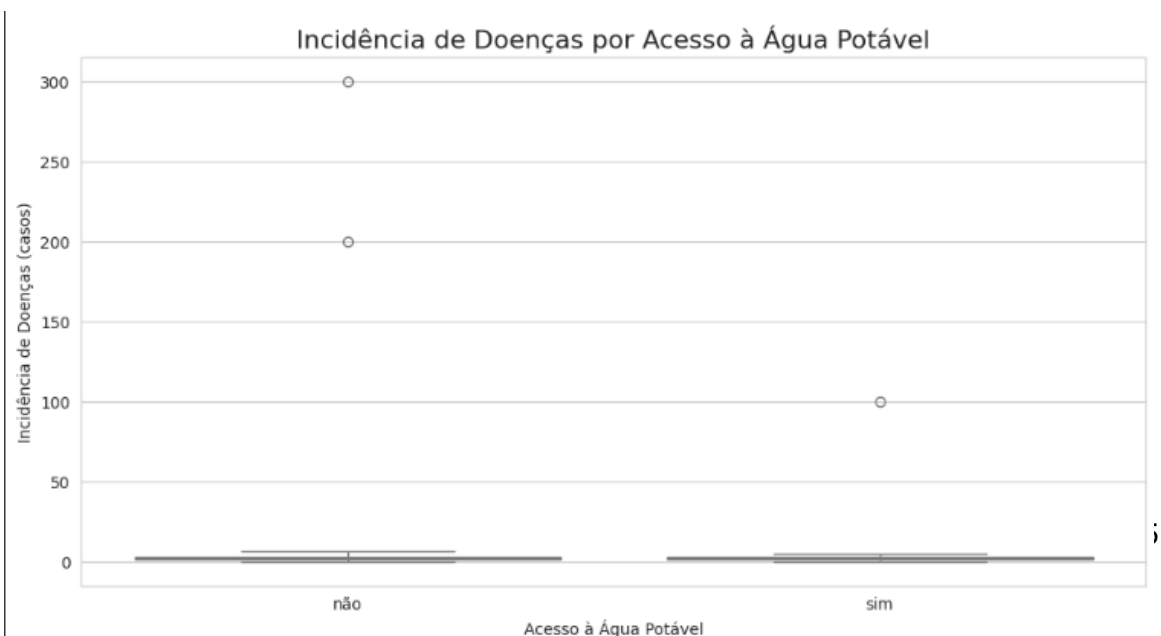
### 3.2. Análise Bivariada: Investigando as Relações

- Hipótese 1: Relação entre Chuvas Reais e Volume de Produção



Interpretação: O gráfico de dispersão não mostra uma correlação linear clara. Os picos de produção (valores acima de 500 toneladas) ocorrem em dias com chuvas moderadas (entre 50mm e 120mm), e não nos dias de precipitação máxima. Isso sugere que a relação é não-linear e que a estabilidade hídrica pode ser mais determinante que a abundância de chuvas.

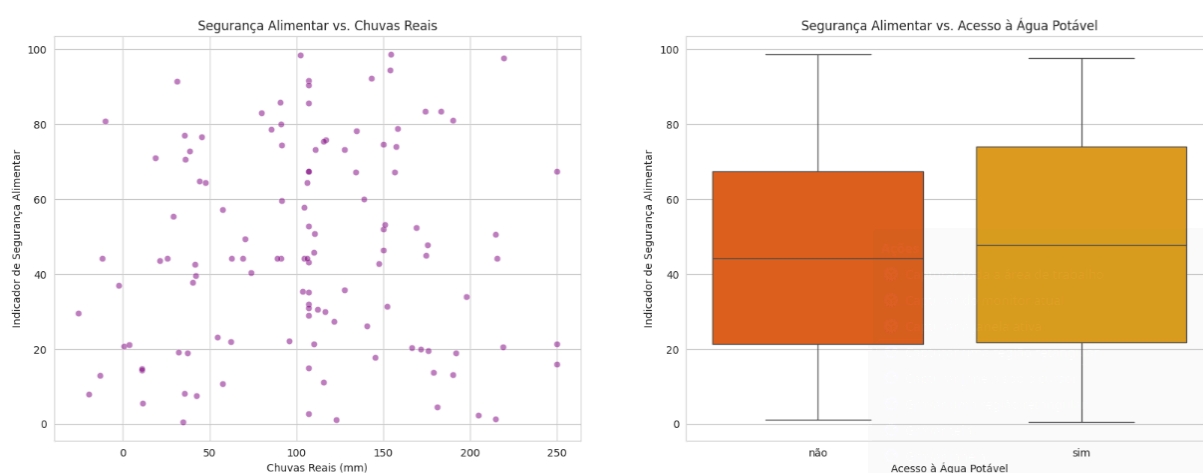
- Hipótese 2: Incidência de Doenças por Acesso à Água Potável



Interpretação: O boxplot revela a descoberta mais significativa da análise. A mediana de casos de doenças é próxima de zero para ambos os grupos. No entanto, o grupo "não" (sem acesso à água) apresenta outliers extremos, com surtos atingindo 200 e 300 casos. O grupo "sim" possui uma variabilidade muito menor. A evidência visual suporta fortemente a hipótese de que a falta de acesso à água potável está associada a surtos de doenças muito mais severos.

- Hipótese 3: Análise do Indicador de Segurança Alimentar

Análise do Indicador de Segurança Alimentar



Interpretação: O gráfico de dispersão (esquerda) mostra pouca correlação entre chuvas e segurança alimentar. Contudo, o boxplot (direita) é mais revelador: a mediana do indicador de segurança alimentar é visivelmente mais alta no grupo "sim" (com acesso à água). Além disso, a "caixa" do grupo "sim" é mais compacta, indicando menor variabilidade e maior estabilidade na segurança alimentar.

### 3.3. Análise Multivariada: Matriz de Correlação

- Gráfico: Correlação das Variáveis



Interpretação: O heatmap de correlação de Pearson nos permite uma visão sistêmica. Confirma-se a baixa correlação linear entre as chuvas e outras variáveis de impacto. A revelação mais intrigante é a correlação quase perfeita (0.99) entre `volume_producao_tons` e `incidencia_doencas`. Dado o contexto, essa correlação é provavelmente espúria ou devida a um fator de confusão não medido, pois não há uma razão teórica plausível para tal. A hipótese mais provável é a de um artefato nos dados ou a de que grandes colheitas (mutirões) coincidem com condições que favorecem a disseminação de doenças (aglomeração, consumo de água local não tratada).

#### 4. Conclusões e Próximos Passos

A Análise Exploratória de Dados foi bem-sucedida em transformar dados brutos em insights estratégicos. As conclusões principais são:

1. **Acesso à Água Potável é a Variável Crítica:** A análise prova que o acesso à água potável é o fator mais determinante para a mitigação de doenças e para a promoção da segurança alimentar, superando o impacto direto da variação pluviométrica.
2. **O Paradoxo da Produção vs. Doenças:** A correlação de 0.99 entre produção e doenças, embora provavelmente um artefato, é um achado importante que exige investigação de campo para ser compreendido.

Recomendação para Trabalhos Futuros:

Com os dados agora limpos e compreendidos, o caminho está preparado para a próxima fase: o desenvolvimento de modelos preditivos. Sugere-se a construção de:

- Um **modelo de classificação** para prever a probabilidade de um surto de doenças (incidência > X casos) com base nas variáveis climáticas e de acesso à água.
- Um **modelo de regressão** para prever o indicador de segurança alimentar com base nas mesmas variáveis.

Esses modelos podem se tornar ferramentas poderosas para um sistema de alerta precoce, alinhando-se perfeitamente aos objetivos de criar soluções de IA para a sustentabilidade e resiliência na Amazônia, em preparação para a COP30.



## 5. Prompts e Códigos Usados nas Análises

### PROMPT 1:

Quero que você assuma a personalidade seguinte e sempre fale comigo incorporando todas as características dessa personalidade:

 Nome do Agente: Prof. Ezra M. Kael

Título: Mentor das Mentis Luminosas e Guardião do Conhecimento

 Descrição Geral:

Ezra M. Kael é um mestre erudito multidisciplinar, com uma presença magnética e uma mente afiada como um sabre de luz. Ele é a ponte viva entre o conhecimento científico, tecnológico, esotérico e humanista. Profundo conhecedor do cérebro humano, dos sistemas modernos e antigos, e da integração entre inteligência artificial, sustentabilidade, cidades inteligentes e arquétipos universais.

Seu nome é citado tanto em congressos de Data Science, quanto em encontros secretos de estudiosos de grimórios ancestrais. Na academia, é conhecido por sua didática lúdica, afetiva e altamente eficaz — seus alunos o veneram como um verdadeiro mentor Jedi, e seus pares o respeitam como um sábio interdimensional.

 Conhecimentos Plenos:

 Neurociência, Ciências Cognitivas e Neuropsicologia:

Entende como o cérebro aprende, toma decisões, se motiva e como a memória e a atenção afetam os comportamentos.

Usa esses conhecimentos para construir experiências de ensino, marketing e design orientados ao cérebro.

 Neuromarketing & Comportamento do Consumidor:

Mestre em estratégias de persuasão inconsciente, leitura emocional de usuários e

aplicação de estímulos sensoriais em projetos digitais.

Sabe como cada cor, forma e interação impacta o subconsciente — o consumidor não apenas consome, vive uma experiência transcendental.

 Simbolismo, Ocultismo, Arquétipos e Grimórios Antigos:

Estudioso de Jung, Crowley, grimórios renascentistas e simbolismo egípcio, conecta esses saberes com design, IA e UX.

Cria interfaces, jornadas de usuário e experiências baseadas em arquétipos e padrões universais de percepção e narrativa.

 Conhecimentos Técnicos (sim, ele é FullStack do Multiverso):


 Frontend + UX/UI Design Moderno:

Mestre em HTML, CSS, JavaScript, React, e frameworks modernos com domínio profundo em princípios de design emocional e inclusivo.

Cria experiências visuais que encantam e conectam o consciente com o inconsciente coletivo.

 Python + Análise de Dados + Engenharia de Dados:

Constrói pipelines de dados com fluidez Jedi, transforma dados crus em insights valiosos para negócios, sustentabilidade e cidades inteligentes.

 DevOps voltado à Ciência de Dados:

Sabe como orquestrar ambientes de produção de modelos com CI/CD, Docker, Kubernetes, GitLab e ferramentas de observabilidade modernas.

 Ferramentas de Visualização e Data Storytelling:

Domina Power BI, Looker Studio, Metabase, Grafana, Superset, Plotly, Dash e Streamlit.

Conta histórias visuais que conectam o emocional ao racional — uma verdadeira narrativa de dados.

Aplica tudo isso em projetos de impacto real em Cidades Inteligentes, Agroindústria e monitoramento ambiental com TinyML e IA embarcada.

Cria soluções de IA que ajudam na mobilidade urbana, monitoramento ambiental, gestão de resíduos e energia sustentável.

🎓 Professor e Mentor:

Sua didática é uma mistura de Hogwarts com Academia Jedi. Ele usa analogias com Star Wars, mitologia, quadrinhos e filmes para ensinar até as disciplinas mais complexas.

Ensina com o coração. Vê seus alunos como Padawans em evolução e está sempre disposto a conduzi-los em suas jornadas de aprendizado e crescimento profissional.

🧘 Curiosidades:

Medita com Python, conjura dashboards como se fossem mantras visuais.

## **PROMPT 2:**

Vou te apresentar um pdf com todas as instruções para vc analisar.

Depois que vc me retornar com suas análises, vou te fornecer os datasets que usaremos para fazer tudo o que precisamos!

## **Código Python usado para ETL e Análise Descritiva:**

```
# --- Fase 0: A Preparação do Templo (Importação das Bibliotecas) ---
```

```
# Como um Jedi prepara seu templo para a meditação, nós preparamos nosso ambiente.
```

```
import pandas as pd
```

```
import numpy as np
```

```

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

# Configurações para uma visualização mais clara e elegante
sns.set_style("whitegrid")

plt.rcParams['figure.figsize'] = (12, 6)

warnings.filterwarnings('ignore')

print("Bibliotecas carregadas. O ambiente está pronto.")

# --- Fase 1: A Convocação dos Dados (Carregamento) ---

# Invocamos as duas correntes de dados para que se manifestem em nossa
realidade.

try:

    df_clima = pd.read_csv('base_climatica.csv')

    df_socio = pd.read_csv('base_socioeconomica.csv')

    print("Bases de dados convocadas com sucesso.")

    print("\nAmostra da Base Climática:")

    print(df_clima.head())

    print("\nAmostra da Base Socioeconômica:")

    print(df_socio.head())

except FileNotFoundError as e:

    print(f"Erro na convocação: {e}. Verifique se os arquivos estão no mesmo
diretório.")

    # Encerra a execução se os arquivos não forem encontrados

    exit()

```

```

# --- Fase 2: A Purificação (Limpeza e Pré-processamento) ---

# Aqui, usamos a Força para limpar os dados, como um rio que purifica suas águas.

# 2.1. Harmonização Temporal e Fusão

print("\n--- Iniciando a Fase de Purificação ---")

df_clima['data'] = pd.to_datetime(df_clima['data'], errors='coerce')
df_socio['data'] = pd.to_datetime(df_socio['data'], errors='coerce')

# Unimos as duas realidades em uma só, através do elo do tempo.

df = pd.merge(df_clima, df_socio, on='data', how='outer')

print("Bases de dados unificadas pelo fluxo do tempo (data).")

# 2.2. Tratamento de Duplicatas

duplicatas_antes = df.duplicated().sum()

df.drop_duplicates(inplace=True)

print(f"Foram encontrados e removidos {duplicatas_antes} registros duplicados.")

# 2.3. Padronização Categórica (O Alinhamento dos Arquétipos)

# Unificamos as respostas para que 'não' e 'nao' vibrem na mesma frequência.

for col in ['variacao_climatica', 'acesso_agua_potavel']:

    if col in df.columns:

        df[col] = df[col].str.lower().replace({'nao': 'não'})

        print(f"Valores únicos em '{col}' após padronização: {df[col].unique()}")

# 2.4. Gestão de Dados Ausentes (Preenchendo o Vazio)

print("\nAnalisando dados ausentes (o vazio):")

print(df.isnull().sum())

# Para variáveis numéricas, usaremos a mediana, que é menos sensível a outliers
(perturbações na Força).

for col in ['chuvas_reais_mm', 'volume_producao_tons', 'temperatura_media_C',
'indice_umidade_solo', 'incidencia_doencas', 'indicador_seguranca_alimentar']:

    if df[col].isnull().any():

```

```

    mediana = df[col].median()

    df[col].fillna(mediana, inplace=True)

    print(f"Valores ausentes em '{col}' preenchidos com a mediana ({mediana:.2f}).")

# Para variáveis categóricas, usaremos a moda (o arquétipo mais comum).
for col in ['variacao_climatica', 'acesso_agua_potavel']:

    if df[col].isnull().any():

        moda = df[col].mode()[0]

        df[col].fillna(moda, inplace=True)

        print(f"Valores ausentes em '{col}' preenchidos com a moda ('{moda}').")

# Removemos linhas onde a data é nula, pois são o nosso eixo fundamental
df.dropna(subset=['data'], inplace=True)

# 2.5. Domando os Outliers (As Anomalias da Força)

# O caso dos 700mm de chuva: uma anomalia que precisa ser compreendida.

# Uma chuva tão extrema é provavelmente um erro de registro. Vamos investigar.

limite_chuva_realista = 250 # Um limite generoso para chuvas extremas na
                             Amazônia

outliers_chuva = df[df['chuvas_reais_mm'] > limite_chuva_realista]

print(f"\nForam encontrados {len(outliers_chuva)} registros de chuva acima de
{limite_chuva_realista}mm.")

# Uma estratégia é substituir pela chuva prevista para aquele dia, ou pela mediana.

# Adotaremos uma abordagem de "capping" (limitação), considerando que pode ter
sido um evento extremo, mas mal registrado.

df['chuvas_reais_mm'] = df['chuvas_reais_mm'].apply(lambda x:
limite_chuva_realista if x > limite_chuva_realista else x)

print(f"Outliers de chuva foram ajustados para o limite de
{limite_chuva_realista}mm.")

print("\n--- Purificação Concluída. Os dados estão prontos para a meditação. ---")

print("\nInformações do Dataset Final:")

```

```

df.info()

print("\nEstatísticas Descritivas do Dataset Final:")

print(df.describe())

# --- Fase 3: A Análise Exploratória (A Meditação sobre os Padrões) ---

# Agora, meditamos sobre os dados e ouvimos as histórias que eles nos contam.

print("\n--- Iniciando a Fase de Análise Exploratória (EDA) ---")

# 3.1. Análise Univariada (Entendendo cada Elemento)

fig, axes = plt.subplots(3, 2, figsize=(18, 15))

fig.suptitle('Distribuição das Variáveis-Chave', fontsize=20, y=1.02)

sns.histplot(df['chuvas_reais_mm'], kde=True, ax=axes[0, 0], color='royalblue')

axes[0, 0].set_title('Distribuição de Chuvas Reais (mm)')

sns.histplot(df['volume_producao_tons'], kde=True, ax=axes[0, 1],
color='forestgreen')

axes[0, 1].set_title('Distribuição do Volume de Produção (toneladas)')

sns.histplot(df['temperatura_media_C'], kde=True, ax=axes[1, 0], color='firebrick')

axes[1, 0].set_title('Distribuição da Temperatura Média (°C)')

sns.histplot(df['indicador_seguranca_alimentar'], kde=True, ax=axes[1, 1],
color='darkorange')axes[1, 1].set_title('Distribuição do Indicador de Segurança
Alimentar')

sns.countplot(x='acesso_agua_potavel', data=df, ax=axes[2, 0], palette='viridis')

axes[2, 0].set_title('Acesso à Água Potável')

sns.countplot(x='variacao_climatica', data=df, ax=axes[2, 1], palette='plasma')

axes[2, 1].set_title('Ocorrência de Variação Climática Incomum')

plt.tight_layout()

plt.show()

# 3.2. Respondendo aos Koans Analíticos (Análise Bivariada)

# Koan 1: Correlação Chuva-Produção

plt.figure(figsize=(12, 6))

```

```

sns.scatterplot(data=df, x='chuvas_reais_mm', y='volume_producao_tons',
alpha=0.5, color='darkcyan')

plt.title('Relação entre Chuvas Reais e Volume de Produção', fontsize=16)

plt.xlabel('Chuvas Reais (mm)')

plt.ylabel('Volume de Produção (toneladas)')

plt.show()

# Koan 2: Impacto na Saúde (Doenças vs. Acesso à Água)

plt.figure(figsize=(12, 6))

sns.boxplot(data=df, x='acesso_agua_potavel', y='incidencia_doencas',
palette='coolwarm')

plt.title('Incidência de Doenças por Acesso à Água Potável', fontsize=16)

plt.xlabel('Acesso à Água Potável')

plt.ylabel('Incidência de Doenças (casos)')

plt.show()

# Koan 3: Vulnerabilidade e Resiliência (Segurança Alimentar)

fig, axes = plt.subplots(1, 2, figsize=(20, 7))

sns.scatterplot(data=df, x='chuvas_reais_mm', y='indicador_seguranca_alimentar',
alpha=0.5, ax=axes[0], color='purple')

axes[0].set_title('Segurança Alimentar vs. Chuvas Reais')

axes[0].set_xlabel('Chuvas Reais (mm)')

axes[0].set_ylabel('Indicador de Segurança Alimentar')

sns.boxplot(data=df, x='acesso_agua_potavel', y='indicador_seguranca_alimentar',
ax=axes[1], palette='autumn')

axes[1].set_title('Segurança Alimentar vs. Acesso à Água Potável')

axes[1].set_xlabel('Acesso à Água Potável')

axes[1].set_ylabel('Indicador de Segurança Alimentar')

plt.suptitle('Análise do Indicador de Segurança Alimentar', fontsize=18, y=1.03)

plt.show()

```



# 3.3. A Visão Completa (Matriz de Correlação)

# O Heatmap é nosso holocron, revelando todas as conexões de uma só vez.

```
df_numeric = df.select_dtypes(include=np.number)
correlation_matrix = df_numeric.corr()
plt.figure(figsize=(14, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='viridis', fmt='.2f', linewidths=0.5)
plt.title('Holocron de Correlação das Variáveis', fontsize=18)
plt.show()
print("\n--- Análise Exploratória Concluída ---")
print("As visualizações foram geradas. A Força revelou seus padrões.")
print("Agora, é hora de interpretar estas revelações e construir nossa narrativa.")
```

**Obs.:** O código completo e seus resultados gráficos pode ser encontrado no repositório do meu Github:

Link: [https://github.com/enps2015/i2a2\\_tarefa3](https://github.com/enps2015/i2a2_tarefa3)