Documentação do Tratamento de Dados

Este documento descreve o processo de tratamento aplicado ao dataset 'dados\_nao\_tratados.csv'. O objetivo foi limpar, padronizar e preparar os dados para análises e futuras cargas em sistemas.

# 1. Carregamento dos Dados

Os dados foram carregados utilizando a biblioteca Pandas do Python, por meio do comando:  
  
df = pd.read\_csv('dados\_nao\_tratados.csv')  
  
Este passo permitiu visualizar as primeiras linhas e identificar problemas como valores nulos, formatação incorreta e inconsistências.

# 2. Tratamento de Valores Nulos e Conversão de Tipos

Foram detectados valores ausentes e formatos inválidos nas colunas 'nota' e 'idade'. As seguintes ações foram realizadas:  
  
- 'nota': convertida para valores numéricos com pd.to\_numeric(errors='coerce'), substituindo valores inválidos por NaN e depois preenchendo com a média da coluna.  
- 'idade': convertida para valores numéricos, substituindo inválidos por NaN e preenchendo com a mediana da coluna.

# 3. Padronização da Coluna 'ativo'

A coluna 'ativo' possuía valores variados como 'Sim', 'sim', 'TRUE', 'Não' e 'e'. Foi aplicada padronização com str.strip().str.lower() e depois mapeamento para valores booleanos:  
- 'sim' ou 'true' → True  
- 'não' ou 'e' → False

# 4. Ajuste das Datas

A coluna 'data\_inscricao' continha diferentes formatos e valores inválidos (ex.: '2022-08-XX'). Foi utilizada a função pd.to\_datetime(errors='coerce') para converter os valores válidos e marcar como NaT (Not a Time) os inválidos, que podem ser removidos ou tratados posteriormente.

# 5. Remoção de Duplicados

Foi aplicado o método df.drop\_duplicates() para eliminar registros repetidos, mantendo apenas a primeira ocorrência.

# 6. Renomeação de Colunas

As colunas foram renomeadas para um padrão uniforme, utilizando df.rename(). Isso facilita futuras análises e padroniza os nomes.

# 7. Exportação dos Dados Tratados

Após os tratamentos, o dataframe final foi exportado para um novo arquivo CSV:  
  
df.to\_csv('dados\_tratados.csv', index=False)  
  
O arquivo 'dados\_tratados.csv' contém os dados prontos para análise e carga em outros sistemas.

# Conclusão

Com as etapas descritas, os dados foram limpos, padronizados e preparados. O processo seguiu as boas práticas de ETL (Extract, Transform, Load), garantindo qualidade e consistência.