Extração e Preparação de Dados

Aula 06 – Dados Ausentes: Identificação e Mecanismos

ibmec

Quem sou eu?

Professor: Luís Aramis dos Reis Pinheiro.

- **Doutorado e Mestrado** em **Ciências Mecânicas** – **UnB – CAPES 7**
- **Graduação** em Licenciatura em Física – **UNIFAP**

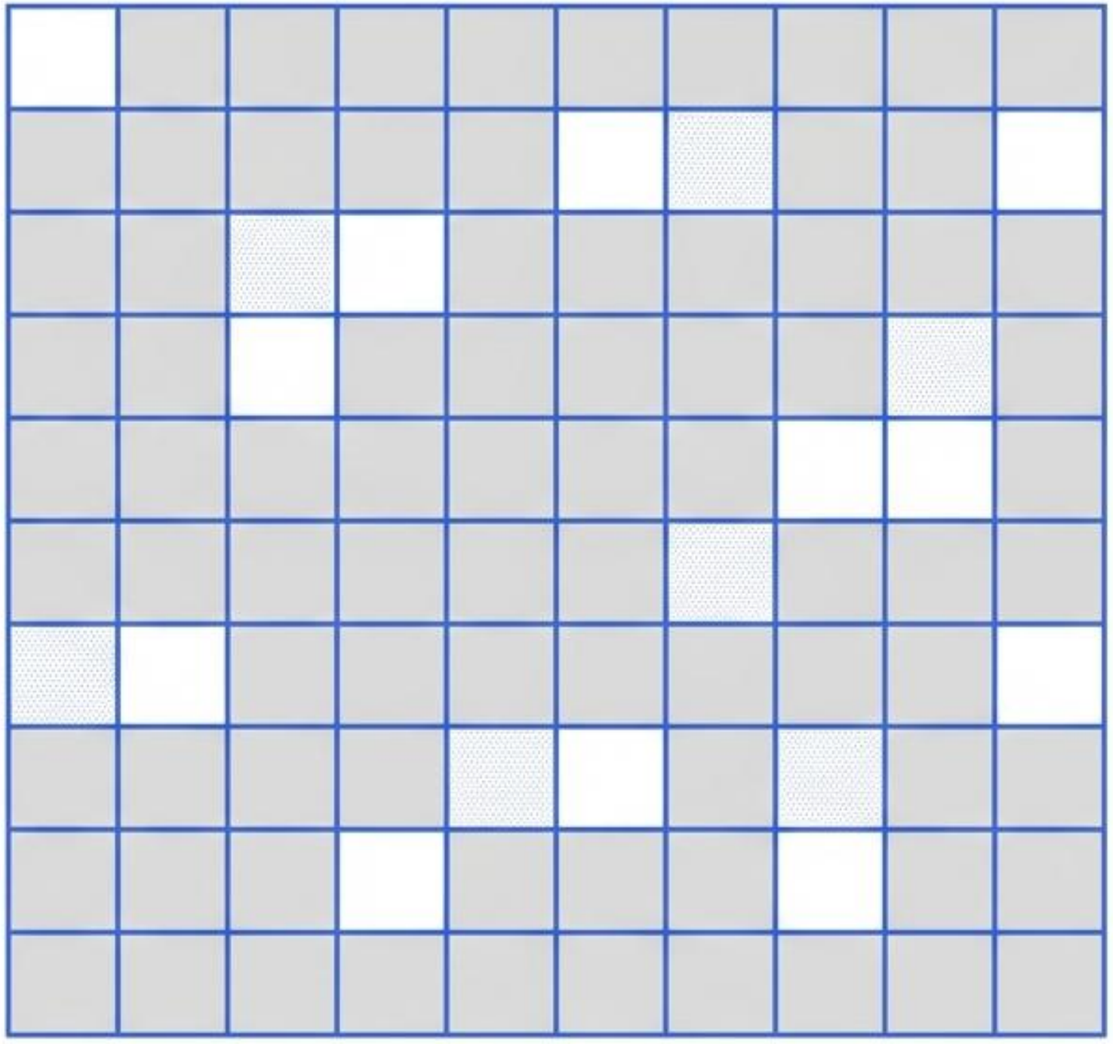
luis.pinheiro@professores.ibemec.edu.br

(96) 99907-5819

@l_aramis

Dados Ausentes: Identificação e Mecanismos

- Compreendendo o "Vazio" antes de Limpar
- Disciplina: Extração e Preparação de Dados (IBM8915)
- Data: 03/03 (Terça-feira)



O Impacto dos Dados Ausentes

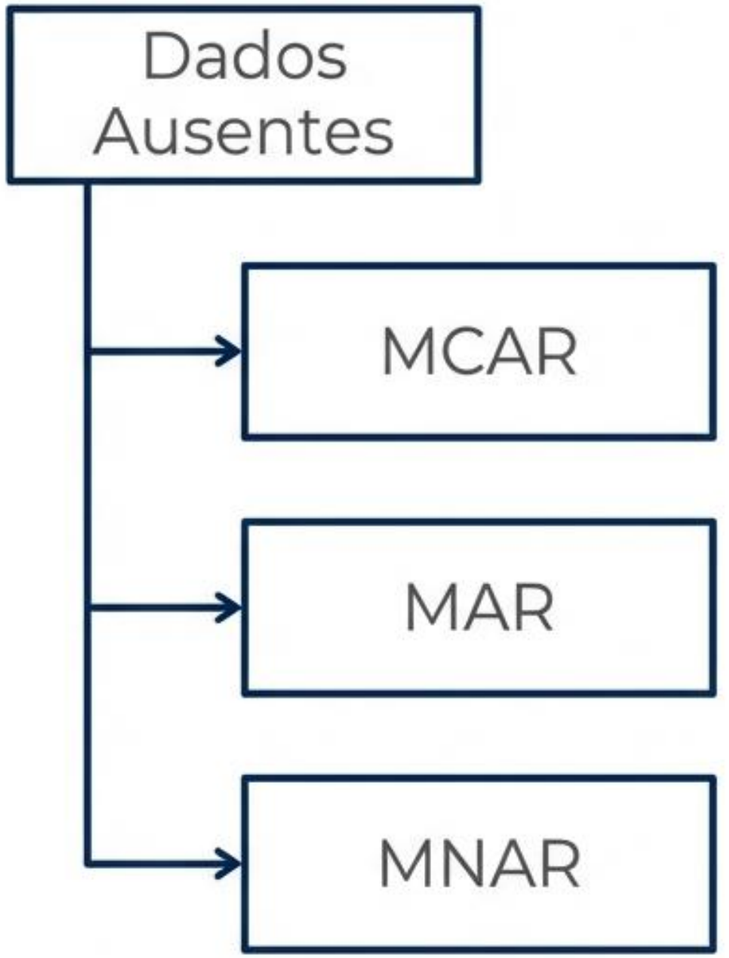
- **A Natureza da Omissão:** Dados reais nunca são 100% completos.
- **Falha Crítica:** Maioria dos algoritmos de Machine Learning não tolera valores nulos.
- **A Armadilha:** Remoção cega introduz viés estatístico letal.

A Teoria de Rubin (Mecanismos)

MCAR: Ausente Completamente ao Acaso (Missing Completely at Random).

MAR: Ausente ao Acaso (Missing at Random).

MNAR: Ausente Não ao Acaso (Missing Not at Random).

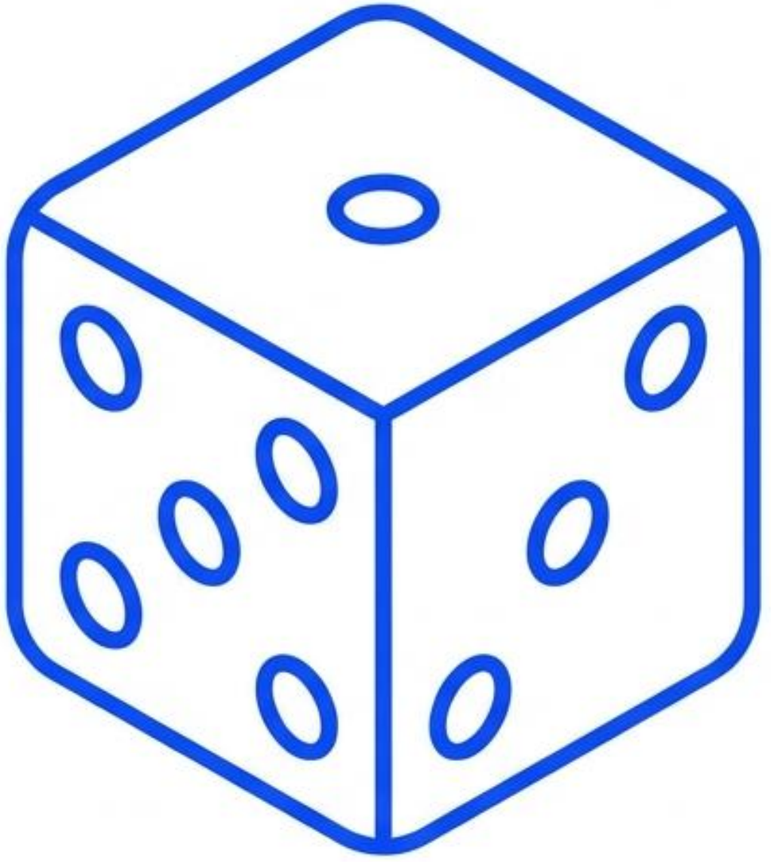


MCAR (Ausente Completamente ao Acaso)

Definição: A omissão independe de qualquer outra variável.

Cenário Ideal: É um evento puramente acidental e aleatório.

Exemplo: Amostra de laboratório destruída por acidente físico.



Check de Aprendizado

Cenário: Pulseira inteligente de monitoramento desliga subitamente por falta de bateria do usuário.

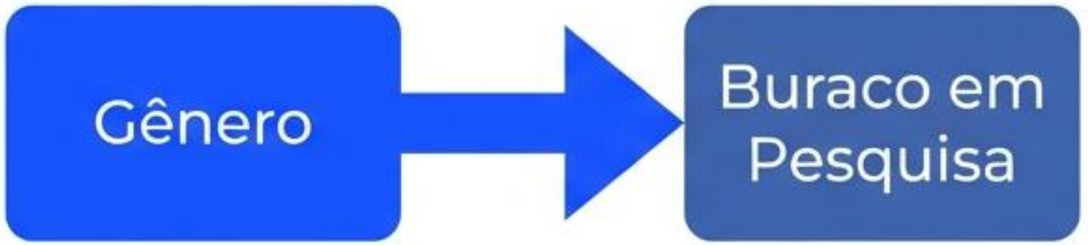
Pergunta: Com base na Teoria de Rubin, este evento é MCAR, MAR ou MNAR?

MAR (Ausente ao Acaso)

Definição: A omissão depende de variáveis observadas no dataset.

Padrão: O dado não sumiu por acidente, mas a causa é explicável.

Exemplo: Homens preenchem menos pesquisas sobre depressão (possuímos o dado de Gênero).



MNAR (Ausente Não ao Acaso)

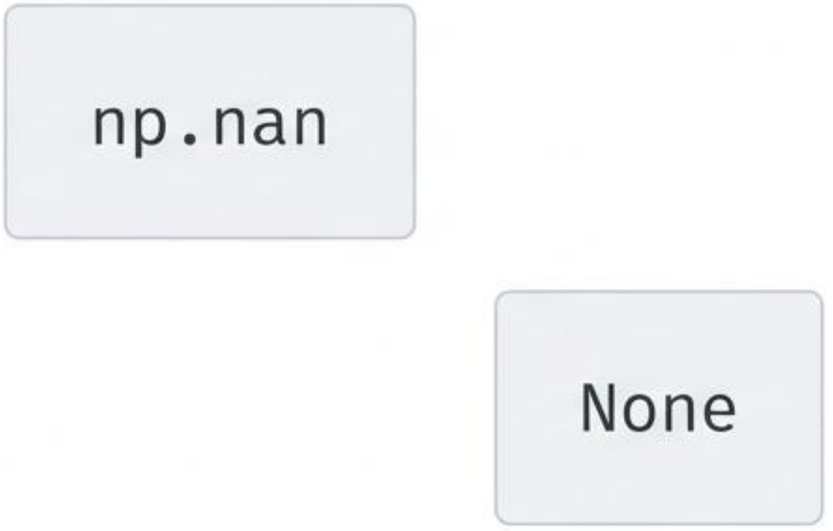
Definição: A omissão depende do próprio valor ausente.

Cenário Crítico: O pior cenário, indica ocultação intencional.

Exemplo: Pessoas de alta renda ocultam seu salário em formulários.

O Valor "Sentinela" no Python

- **Desafio Técnico**: Como a máquina processa o "nada"?
- **Solução**: Injeção de Valores de Sentinela.
- **NaN (Not a Number)**: Padrão (NumPy) para floats.
- **None**: Objeto nativo do Python para strings.

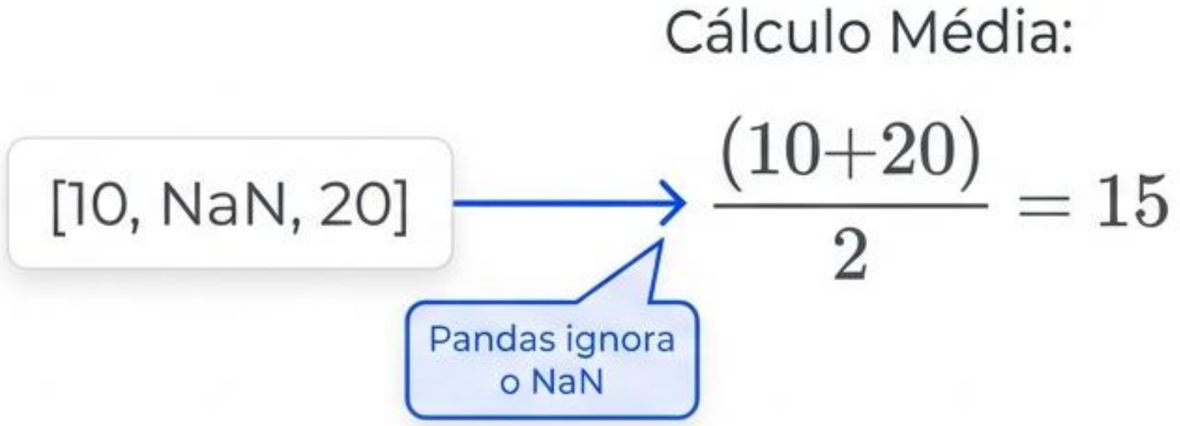


O Comportamento Silencioso do Pandas

Identificação: Reconhecimento nativo de NaN e None.

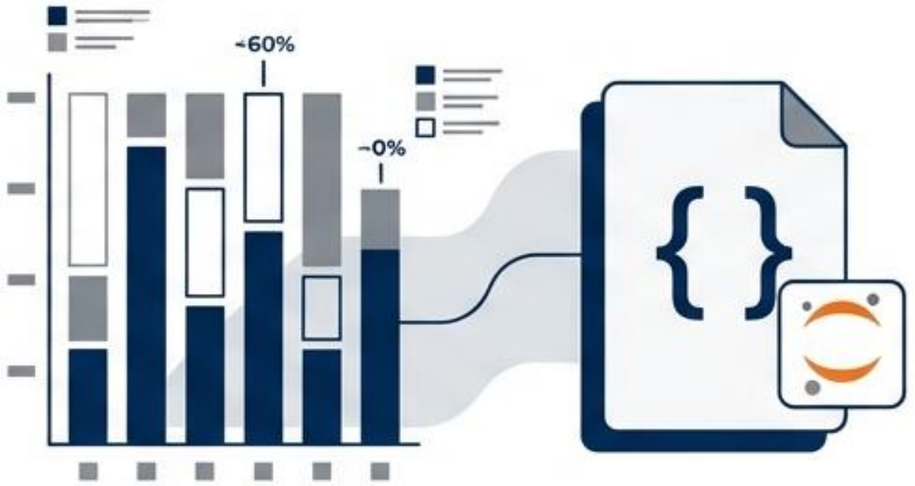
Exclusão Automática: Métodos como **.mean()** ou **.describe()** ignoram nulos.

Atenção: Facilita o cálculo estatístico, mas mascara o problema estrutural.



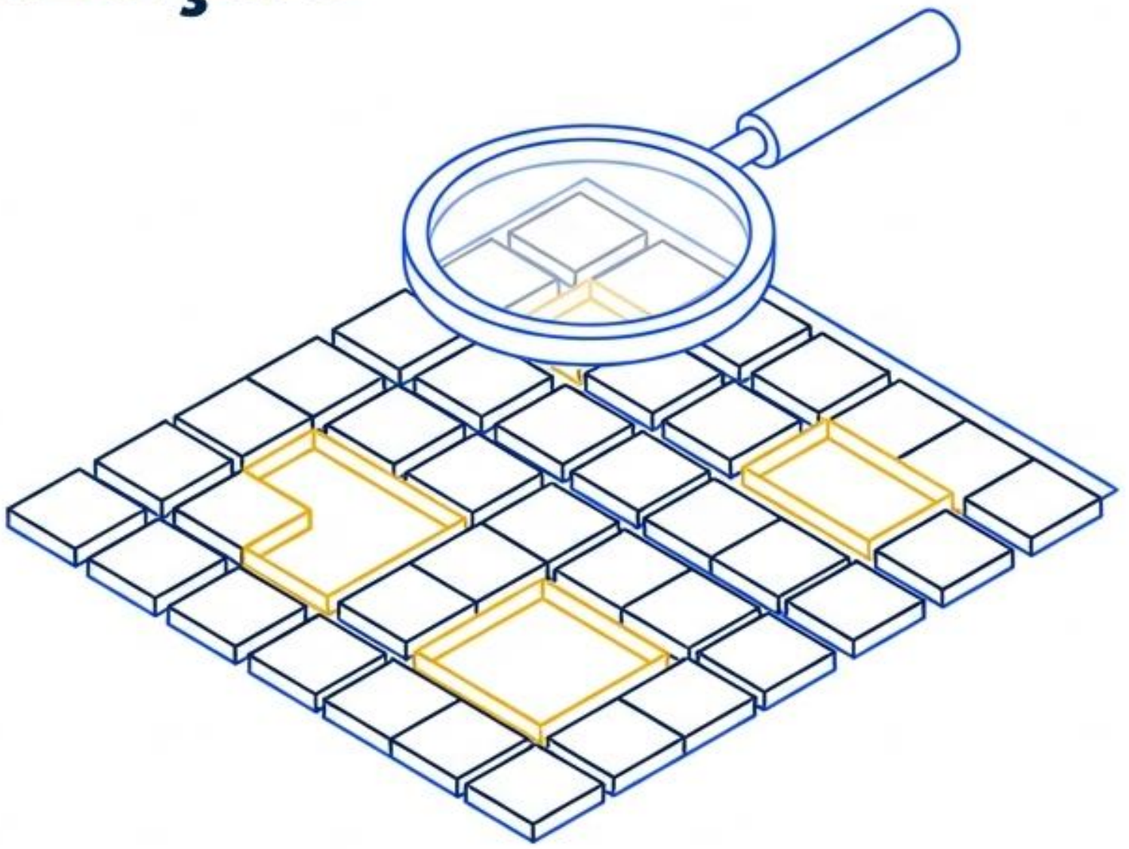
Ponte para a Prática (Laboratório)

- Do Conceito à Caça aos Nulos.
- Quantificação: Uso de df.isnull().sum().
- O Artefato Visual: Construção do "Raio-X" da completude do Dataset.



Diagnóstico Prático: Mapeamento e Visualização

- Transição da teoria (Mecanismos de Ausência) para a prática (Código).
- O uso do Python para rastrear e quantificar valores sentinela.
- Construção do "Raio-X" de completude do dataset.



Mapeamento Quantitativo de Nulos

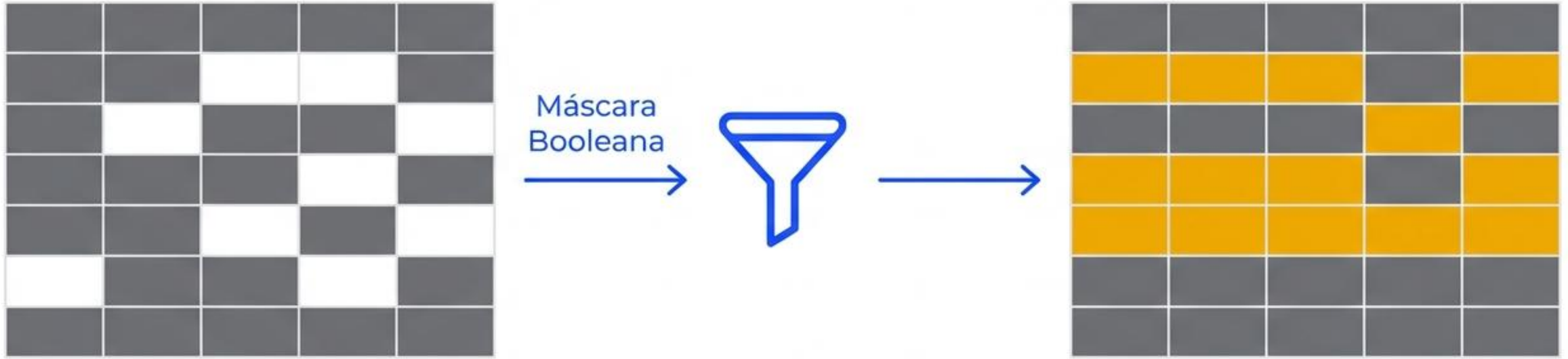
- A Detecção: O método df.isnull() varre o DataFrame em busca de NaN.
- A Quantificação: O encadeamento .sum() consolida a contagem por coluna.
- O Diagnóstico Rápido: Identificação imediata das variáveis mais críticas.

```
df.isnull().sum()
```

```
PassageiroId:  0
Sobreviveu:    0
Classe:        0
Nome:          0
Sexo:          0
Idade: 177
Bilhete:       0
Tarifa:        0
Cabine: 687
```

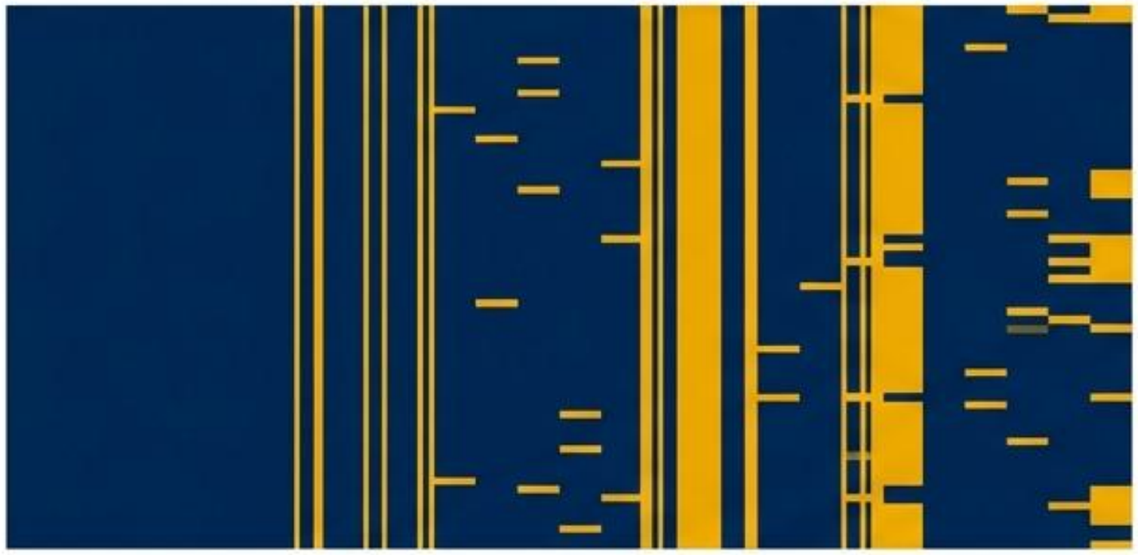
Isolando o Problema com Máscaras Booleanas

- Filtragem avançada: Extraindo apenas as linhas defeituosas.
- Sintaxe: `df[df['coluna'].isnull()]`
- Uso analítico: Buscar padrões de ausência simultânea (mecanismos MAR/MNAR).



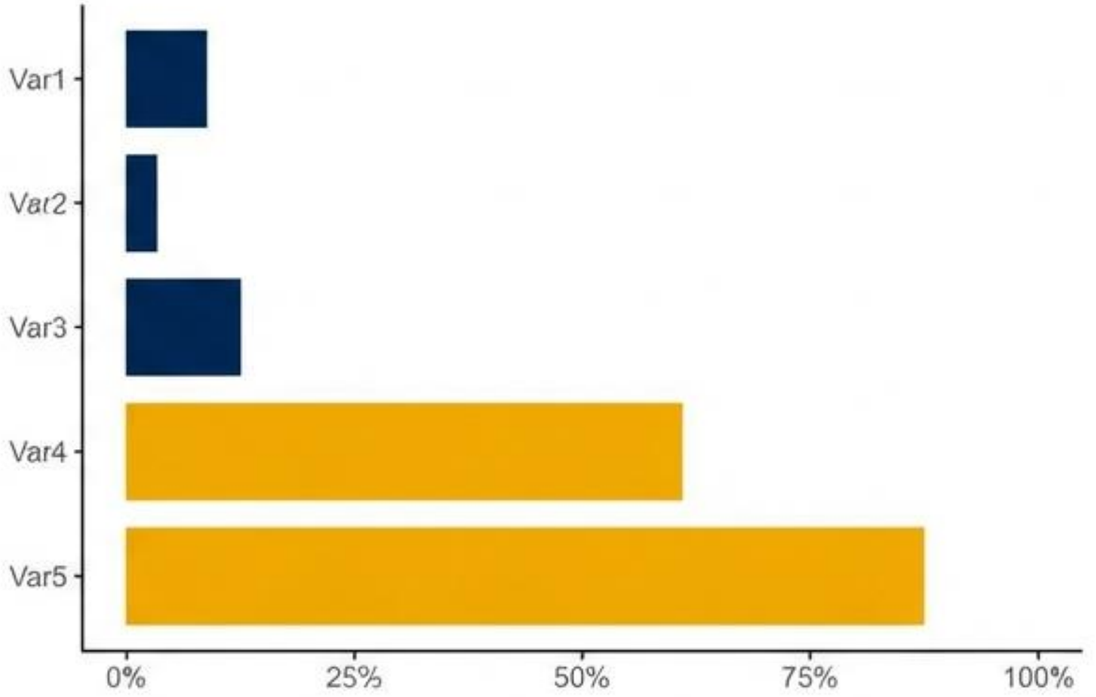
O Raio-X Visual: Seaborn Heatmap

- Gráficos matriciais para visualização de completude.
- Integração direta:
  `sns.heatmap(df.isnull(), cbar=False)`
- Identificação imediata de padrões de "buracos" no dataset.



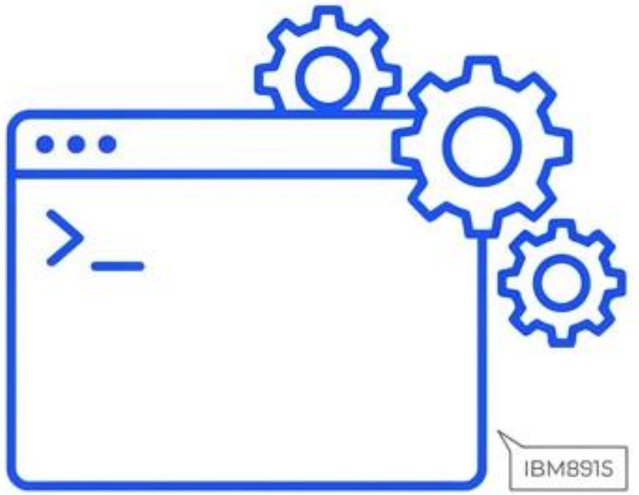
Proporção de Ausência: Gráfico de Barras

- Transformando volume absoluto em impacto percentual.
- Cálculo: `(df.isnull().sum() / len(df)) * 100`
- Plotagem ordenada para relatórios de Qualidade de Dados.



Mão na Massa: Laboratório de Visualização

- Ambiente: Jupyter Notebook.
- Arquivo: lab_04_missing_values_viz.ipynb
- Missão: Diagnosticar, calcular percentuais e renderizar o Heatmap.



Avaliação e Versionamento (Portfólio)

— O código só existe se estiver versionado.

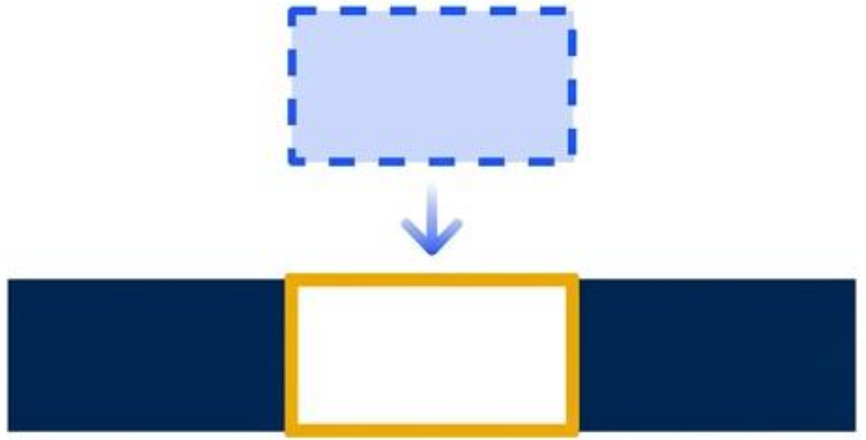
```
git add lab_04_missing_values_viz.ipynb

git commit -m "feat: analise e heatmap de dados ausentes"

git push origin main
```

Próxima Etapa:
A "Cirurgia" dos Dados

- Aula 07 (Quinta-feira): Tratamento Univariado.
- Transição do Diagnóstico para a Imputação.
- Introdução ao método fillna().
- Uso estatístico: Média, Mediana e Moda.



Dúvidas?


CRIATIVIDADE
INOVAÇÃO
¡DEIA!
COMAMUNIDADE
CRIATIVIDADE
BRAINSTORM
INOVAÇÃO

Ibmec

A direção do amanhã.

/ibmec