Extração e Preparação de Dados

Aula 05 – Qualidade de Dados e Metadados

IBEMEC - 2026.1 | Prof. Luís Aramis

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**

aramisrp@gmail.com

(96) 99907-5819

@l_aramis

Qualidade de Dados e Metadados

Avaliação de Saúde do Dado & Construção de Dicionários

Disciplina: Extração e Preparação de Dados (IBM8915)
Prof. Luís Aramis | 26/02

O Princípio GIGO (Garbage In, Garbage Out)

Premissa:
Todos os dados do mundo real são 'sujos'.

Objetivo:
Tornar os dados menos sujos, não perfeitos.

Impacto:
Dados ruins > Processo Perfeito > Decisão Ruim.

Foco:
Qualidade da Coleta vs. Tratamento.

Dimensões da Qualidade de Dados

Acurácia

O dado reflete a realidade? (Factualidade).

Completude

Existem valores ausentes? (Nulos).

Tempestividade

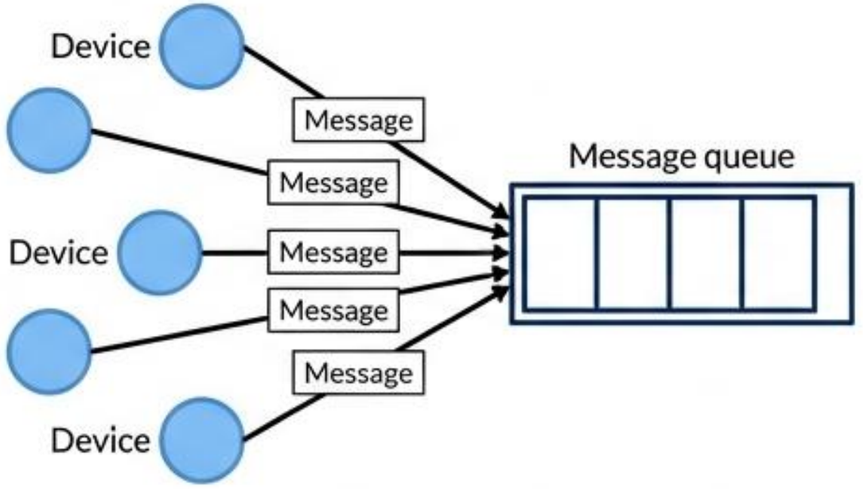
O dado está disponível a tempo? (Latência).

Qualidade Técnica vs. Valor de Negócio

Qualidade

Atributos técnicos (ex: preenchimento 100%).

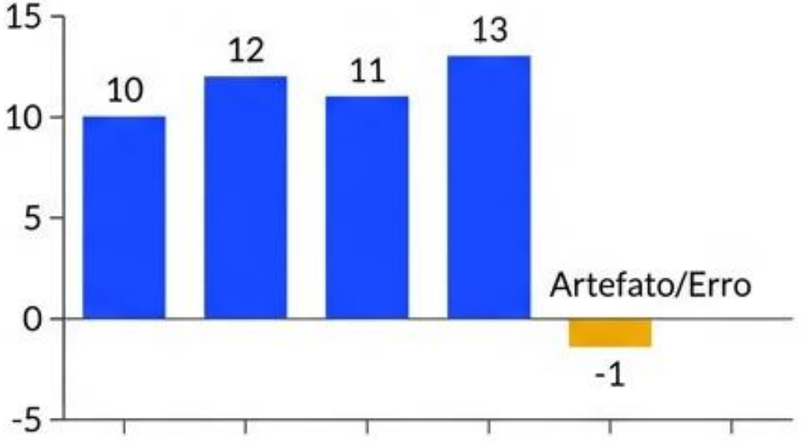
Artefatos de Dados: Erros sistêmicos não intencionais.



Utilidade

Capacidade de responder à pergunta de negócio.

Causa Raiz: Falhas de sensores, software ou transcrição.



Desafio Rápido: O Caso das Duplicatas

Cenário: Uma falha no sistema duplicou todas as vendas de sexta-feira.

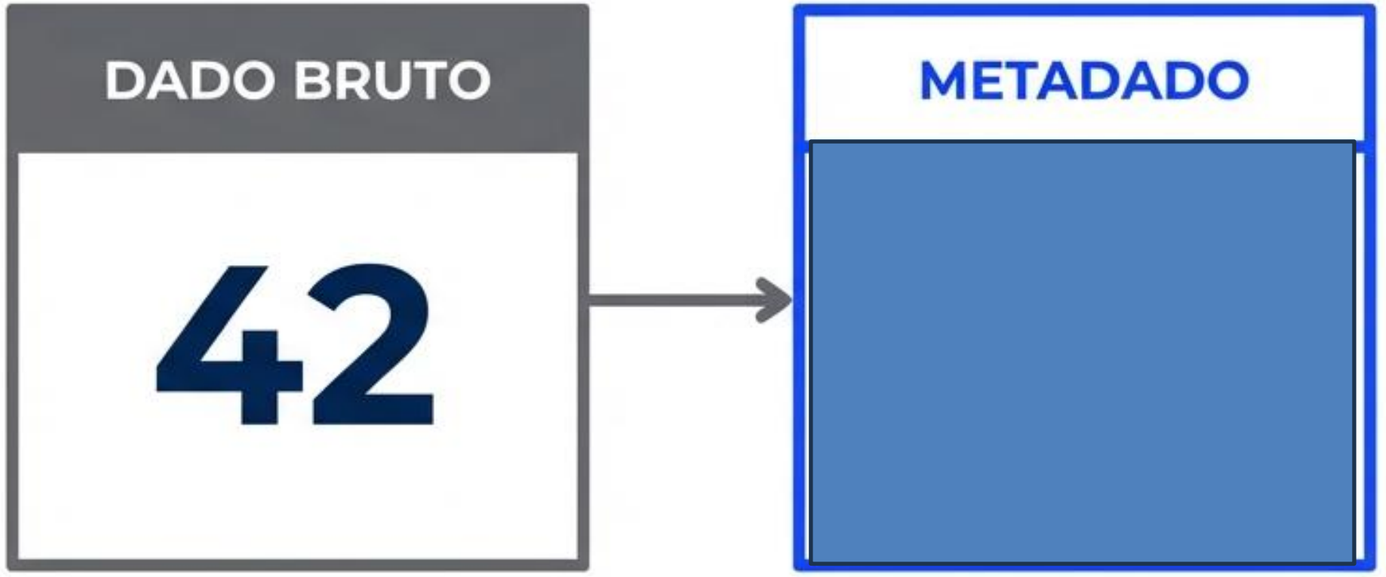
Pergunta: Qual pilar da qualidade foi comprometido?

A) Completude

B) Tempestividade

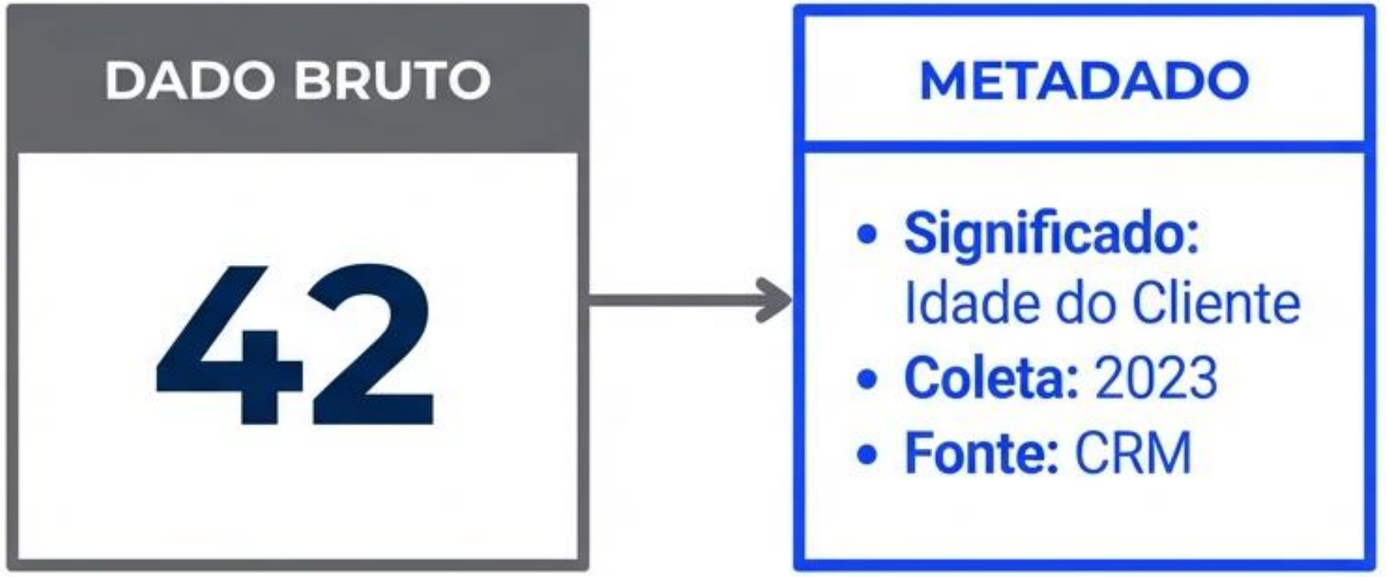
C) Acurácia

A Fundação da Governança: Metadados



- **Definição:** Dados sobre os dados.
- **Função:** Contexto, Rastreabilidade e Auditoria.
- **Objetivo:** Transformar números brutos em informação.
- **Analogia:** O rótulo nutricional de um produto.

A Fundação da Governança: Metadados



- **Definição:** Dados sobre os dados.
- **Função:** Contexto, Rastreabilidade e Auditoria.
- **Objetivo:** Transformar números brutos em informação.
- **Analogia:** O rótulo nutricional de um produto.

Classificação dos Metadados

Negócios

Regras e definições lógicas (O que é?)

Técnicos

Schemas, tipos e formatos (Como é?)

Operacionais

Logs de execução e status (Quando foi?)

Referência

Tabelas de domínio e lookups (De-Para)

Classificação dos Metadados

Negócios

'O que é um cliente?

Técnicos

O CPF é do tipo String ou Int64?

Operacionais

O script de extração falhou às 03:00 da manhã

Referência

E os de Referência são seus dados de 'de-para' ou Lookups, como a tabela de CEPs

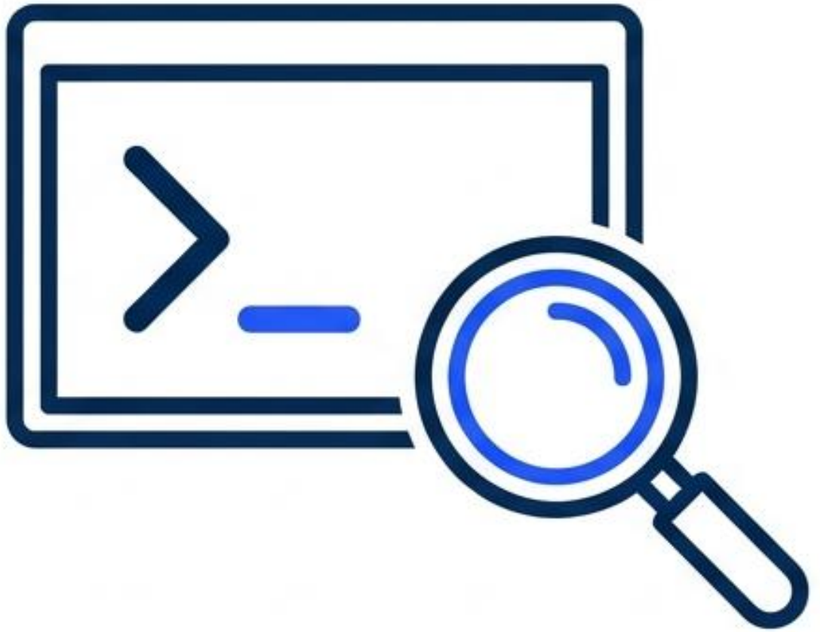
O Artefato: Dicionário de Dados

- **Formato**: Documento Markdown (.md).
- **Variável**: Nome da coluna no banco.
- **Tipo**: Classificação técnica (int64, object, float).
- **Descrição**: Significado e regras de negócio.

```
| Variável   | Tipo   | Descrição |
| :---       | :---   | :---
| user_id    | int64  | Identificador único (PK) |
| churn_flag | int64  | 1 = Cancelou, 0 = Ativo |
| trans_date | object | Data da transação (ISO8601) |
```

Missão: O Auditor de Dados

- **Atividade:** Diagnóstico de um Dataset 'Sujo'.
- **Ferramentas:** `isnull()`, `duplicated()`, `value_counts()`.
- **Entregável:** `data_dictionary.md` (GitHub).
- **Objetivo:** Mapear anomalias antes de corrigir.



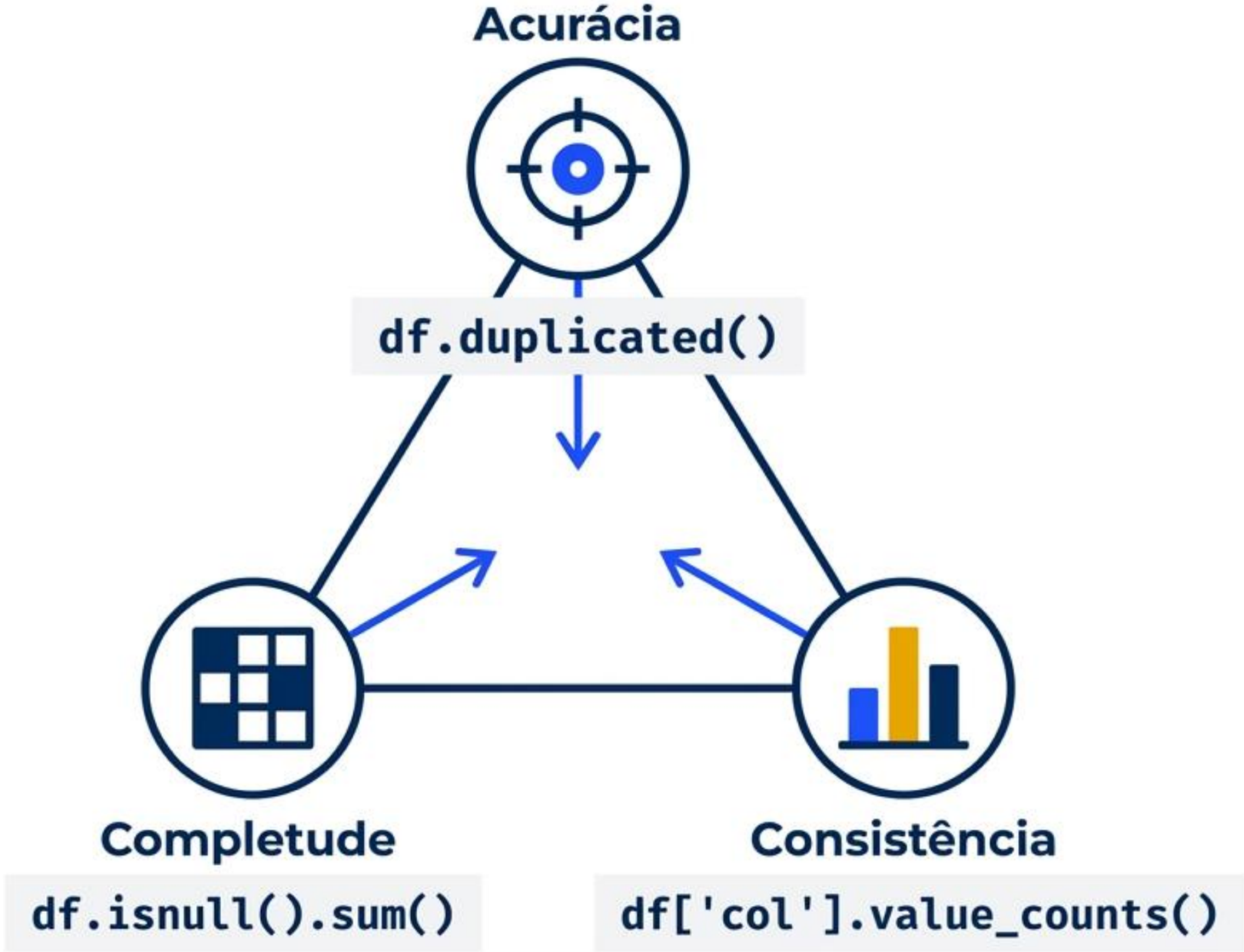
Diagnóstico de Qualidade de Dados: Prática

Aula 05 (Parte 2) | Extração e Preparação de Dados

Do Conceito à Auditoria Técnica

Teoria (Concluído) → Ferramental Pandas → Laboratório PBL → Entrega

Ferramentas de Diagnóstico Clínico no Pandas



Acurácia: Detecção de Redundância

```
# Contagem absoluta
df.duplicated().sum()

# Visualização dos registros
df[df.duplicated()]
```

- Distinção crítica: Duplicata de **ID (chave)** vs. Linha completa.

	ID	Nome	Valor	Data
1	101	Alice	500	2023-10-27
2	102	Bob	750	2023-10-28
3	103	Charlie	600	2023-10-29
4	102	Bob	750	2023-10-28

Duplicata

Completude: Mapeamento de Lacunas

```
# Absoluto vs Relativo
nulos = df.isnull().sum()
percentual = (nulos / len(df)) * 100
```

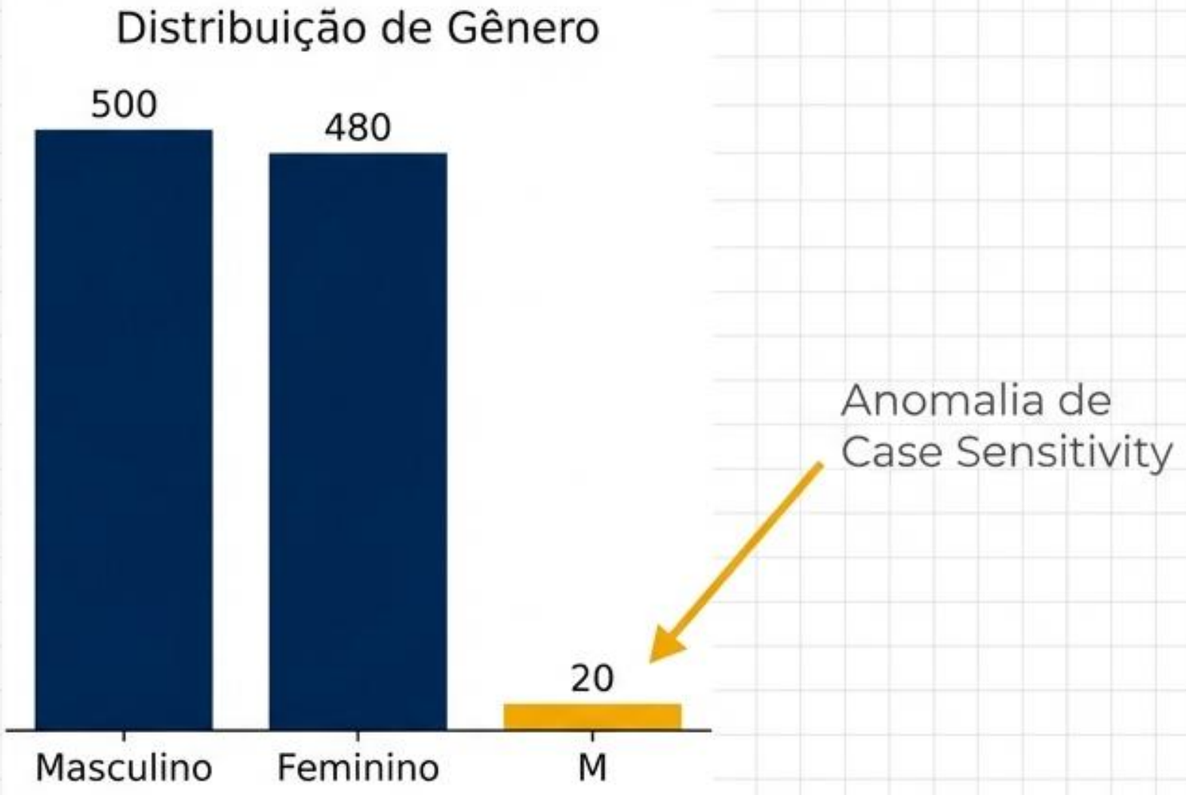
```
ID              0       (0.00%)
Salário         13      (15.00%)
Comentários     78      (90.00%)
```

O número absoluto engana.
Foque na proporção.

Consistência: Análise de Frequência e Domínio

```
df['genero'].value_counts()
```

```
Masculino    500
Feminino     480
M             20
```



Laboratório PBL: “O Dataset Sujo”

Missão: Auditoria do Dataset RH_Corrupted.csv

- [ ] 1. Carregar e Inspecionar (info, head)
- [ ] 2. Quantificar Duplicatas
- [ ] 3. Calcular % de Nulos em 'Salário'
- [ ] 4. Identificar anomalias em 'Departamento'

Execução: Carga e Acurácia

Carga

```
df = pd.read_csv('RH_Corrupted.csv')
```

→ Verificar Dtype

Um ID duplicado com dados diferentes é erro ou atualização de cadastro?

Acurácia

```
df.duplicated(subset=['ID']).sum()
```


```
df.duplicated(subset=['ID']).sum()
```


```
In [ ]:
```

Execução: Completude e Domínio

Foco na coluna Salário

```
df['salario'].isnull().mean()
```

Foco na coluna Idade

```
df[df['idade'] < 0]
```

```
# Idades negativas são
impossíveis. Anotar.
```

O Artefato: Dicionário de Dados

Variável (Nome Técnico)

`data_nascimento`

Tipo (Int, Float, Obj)

`Datetime64[ns]`

Descrição (Significado de Negócio)

Data de nascimento do cliente, formatada como AAAA-MM-DD.

Problemas Conhecidos (Diagnóstico)

Algumas datas estão anteriores a 1900, indicando erro de preenchimento ou dados legados.

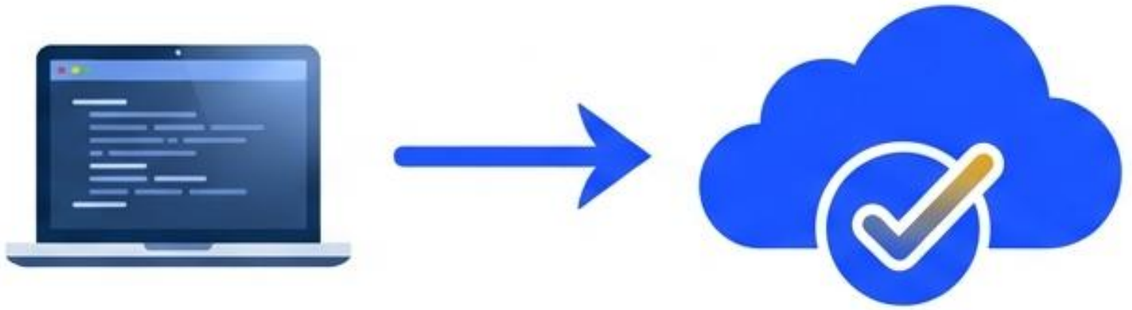
Documento vivo que traduz o 'tecniquês' para a linguagem de negócios.

Estruturando o data_dictionary.md

Variável	Tipo	Descrição	% Nulos	Anomalias Identificadas
id_func	Int	Identificador único	0%	**2 IDs duplicados**
salario	Float	Salário bruto mensal	15%	-
uf	Obj	Estado de residência	0%	**'SP' e 'Sao Paulo' misturados**

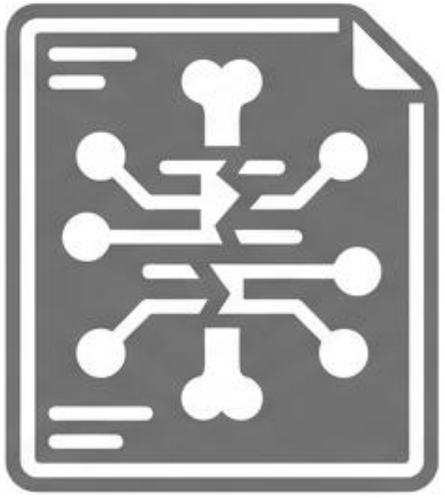
Consolidação no Portfólio (Git)

```
git add data_dictionary.md

git commit -m "feat: adiciona
dicionario de dados e auditoria
inicial"

git push origin main
```



Critério de Sucesso: Arquivo visível no repositório remoto.

Próxima Aula: A Cirurgia



HOJE: Diagnóstico

Identificamos lacunas e inconsistências na estrutura dos dados, como valores ausentes e formatos inválidos, similar a fraturas ósseas ou doenças visíveis nos exame.

PRÓXIMA AULA: Tratamento

Aula 06 e 07: Imputação de Dados e Remoção de Outliers. Aplicaremos técnicas precisas precisas para corrigir os dados, preencher lacunas e eliminar anomalias, restaurando a integridade do dataset.

Hoje diagnosticamos. Amanhã operamos.

Dúvidas?


CRIATIVIDADE
INOVAÇÃO
¡DEIA!
COMAMUNIDADE
CRIATIVIDADE
BRAINSTORM
INOVAÇÃO

Ibmec

A direção do amanhã.

/ibmec