Etapa 3 do CRISP-DM - Limpeza e Tratamento de Dados (Preparação dos Dados)

Para cada dataset que a equipe tiver, deve-se fazer:

1. Entendimento inicial dos dados

- Nome do dataset:
- Carregar o dataset e verificar dimensões (linhas × colunas).
- Visualizar as primeiras linhas para ter ideia da estrutura.
- Obter informações das variáveis (tipo de dados, categorias, numéricas, datas, etc.).
- Estatísticas descritivas (média, mediana, desvio padrão, mínimo, máximo, contagem).

2. Tratamento de valores ausentes (missing values)

- Identificar variáveis com valores nulos ou ausentes.
- Decidir estratégia de tratamento:
    - Remover registros incompletos (quando a perda for pequena).
    - Imputar valores (média, mediana, moda, valor constante, interpolação).

3. Tratamento de inconsistências e duplicatas

- Verificar e remover linhas duplicadas.
- Conferir valores inconsistentes (ex.: idade = -5, salário = 9999999, datas impossíveis).
- Padronizar formatos (ex.: datas no mesmo padrão, strings em minúsculo/maiúsculo uniforme).

4. Padronização e normalização dos dados
Uniformizar categorias (ex.: “SP”, “sp”, “São Paulo” → “SÃO PAULO”).

5. Detecção e tratamento de outliers

- Identificar outliers em variáveis numéricas (usando boxplot ou outra ferramenta).

- Decidir tratamento (e justificar decisão):
    - Remover registros com outliers extremos.
    - Substituir valores por limites aceitáveis (winsorização).
    - Manter, se fizer sentido para o contexto.

---

6. Criação de variáveis (Feature Engineering)

- Criar novas variáveis úteis (ex.: idade a partir da data de nascimento, mês/ano a partir de datas).
- Transformar variáveis categóricas em dummies ou embeddings.
- Agrupar variáveis pouco representativas em categorias mais amplas.

8. Documentação e salvamento

- Registrar todas as transformações feitas (para garantir **reprodutibilidade**).
- Salvar o dataset limpo em um novo arquivo (CSV/Parquet/etc.).
- Manter o dataset original intacto.

Entrega:

- Escrever brevemente **a justificativa das escolhas** (ex.: “removi 5 linhas duplicadas”, “substituí valores nulos de renda pela mediana”).