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

**1. Dataset: olist\_orders\_dataset\_R01.csv**

**Descrição:**  
Contém informações sobre os pedidos, como status, datas de compra e entrega.

**Tratamento realizado:**

* Conversão das colunas de data para formato datetime.
* Preenchimento de valores ausentes com datas estimadas ou de compra.
* Remoção de duplicatas e padronização de texto no campo order\_status.
* Criação das variáveis delivery\_delay\_days e delivered\_on\_time.

**Resultado:**  
99.441 registros válidos.

**2. Dataset: olist\_order\_items\_dataset\_R01.csv**

**Descrição:**  
Contém os itens de cada pedido, com preço e frete.

**Tratamento realizado:**

* Tipos numéricos em price e freight\_value.
* Remoção de duplicatas e registros com order\_item\_id inválido.
* Criaço de variável total\_item\_value = price + freight\_value.

**Resultado:**  
112.650 itens de pedidos, limpos e prontos para cálculo de custos e frete médio.

**3. Dataset: olist\_order\_payments\_dataset\_R01.csv**

**Descrição:**  
Registra os tipos e valores de pagamento por pedido.

**Tratamento realizado:**

* Conversão de colunas numéricas (installments, value).
* Padronização de payment\_type (minúsculo, sem espaços).
* Criação de variável is\_installment para indicar pagamentos parcelados.

**Resultado:**  
103.886 registros de pagamento válidos

**4. Dataset: olist\_order\_reviews\_dataset\_R01.csv**

**Descrição:**  
Contém avaliações de clientes, notas e comentários.

**Tratamento realizado:**

* Conversão de datas de criação e resposta para datetime.
* Preenchimento de comentários nulos com texto vazio.
* Padronização de IDs e remoção de duplicatas.
* Criação de review\_sentiment (positiva, neutra, negativa).

**Resultado:**  
+-100.000 avaliações válidas.

**5. Dataset: olist\_products\_dataset\_R01.csv**

**Descrição:**  
Reúne dados sobre dimensões e categorias de produtos.

**Tratamento realizado:**

* Preenchimento de medidas físicas ausentes com mediana.
* Padronização de categorias (minúsculo, sem espaços).
* Junção com tabela de tradução (product\_category\_name\_translation).
* Criação de product\_volume\_cm3 = comprimento × altura × largura.

**Resultado:**  
32.951 produtos válidos

**6. Dataset: olist\_sellers\_dataset\_R01.csv**

**Descrição:**  
Informações cadastrais de vendedores.

**Tratamento realizado:**

* Remoção de duplicatas.
* Padronização de seller\_city e seller\_state (maiúsculo, sem espaços).
* Criação de seller\_region\_code (estado + prefixo de CEP).

**Resultado:**  
3.095 vendedores padronizados para mapeamento logístico por região.

**7. Dataset: olist\_geolocation\_dataset\_R01.csv**

**Descrição:**  
Base de geolocalização de CEPs com latitude e longitude.

**Tratamento realizado:**

* Remoção de registros nulos em coordenadas.
* Padronização de cidade e estado em maiúsculo.
* Criação de geo\_region\_code (estado + prefixo de CEP).

**Resultado:**  
100.016 pontos geográficos válidos.

**8. Dataset: product\_category\_name\_translation\_R01.csv**

**Descrição:**  
Tabela de referência para tradução de categorias de produto.

**Tratamento realizado:**

* Remoção de linhas nulas e duplicadas.
* Padronização de texto.

**Resultado:**  
71 alterações ortográficas.