Registro de Inconsistências, Anomalias e Transformações no Dados

Status	Concluído

🔍 Orientações Gerais

- Os dados de Data devem estar no formato: AAAA-MM-DD (sem horários)

📋 Tabela: olist_geolocation

- Linhas duplicadas removidas
- Coluna **geolocation_zip_code_prefix**: Identificada como "Object", deve ser "INT"
- Coluna **geolocation_city**: 8011 valores únicos (Brasil possui 5570 municípios) - Erro de padronização com variações por acento, caixa e espaço (-2044 registros)
- Coluna **geolocation_city**: Presença de abreviações (sp, rj, bh...), e erros de digitação.

👤 Tabela: olist_customers

- Coluna **customer_zip_code_prefix**: Identificada como "Object", deve ser "INT"

🏪 Tabela: olist_sellers

- Coluna **seller_zip_code_prefix:** Identificada como "Object", deve ser "INT"
- Coluna **seller_city**: Erro na padronização/digitação dos dados

- Coluna **seller_city**: Presença de abreviações (sp, sbc)
- Coluna **seller_city**: Valor "04482255" (com base no CEP, foi substituído por "rio de janeiro")

📦 Tabela: olist_products

- Colunas com dados faltantes (610 linhas): **product_category_name** → Solução: Substituído por "Não Definido"
- Para análise, será usado apenas as colunas **product_id** e **product_category_name**

📝 Tabela: olist_orders

- Pedidos com **order_status = "delivered"** que possuem datas nulas devem ser excluídos.

  Observação: Valores nulos em colunas de datas com outros tipos de order_status podem ser aceitáveis
- Coluna **order_status**: Traduzir os valores

📦 Tabela: olist_order_items

- Coluna **shipping_limit_date**: 4 registros com datas de 2020 - **Removidos**

💰 Tabela: olist_order_payments

- Coluna **payment_installments**: Registros com valor zero - imputar pela média
- Coluna **payment_value**: Registros com valor zero - imputar pela média
- Coluna **payment_type**: Substituir valores "not_defined" pela moda

⭐ Tabela: olist_order_reviews

- Comentários consolidados em uma nova coluna "full_review_text"

  Explicação: *Observou-se que muitos usuários inserem seus comentários incorretamente no campo “review_comment_title”, deixando a coluna*

“review_comment_message” vazia. Em outros casos, o oposto ocorre: o comentário é registrado apenas na mensagem, sem título, resultando em valores nulos na coluna de título.

Para facilitar a análise e garantir que nenhum conteúdo relevante seja perdido, foi criado o DataFrame ***“comentarios”****, com destaque para a nova coluna* ***“full_review_text”****, que consolida o título e a mensagem em um único campo. Dessa forma, todos os comentários ficam completos e livres de valores nulos.*