Verificar a Qualidade dos Dados

1) Diagnóstico (resumo)

- **Missing (ausência de dados)**
    - **Order:** ~4,67% global; altos em `extraInfo` (59,05%) e `scheduledAt` (48,40%).
    - **Customer:** ~11,73% global; altos em `externalCode` (62,60%), `enrichedAt`/`enrichedBy` (49,60%), `gender` (25,90%).
    - **CampaignQueue:** ~6,41% global; altos em `response` (68,24%) e `sendAt` (34,24%).
    - **Campaign:** ~2,59% global; `description` (20,25%), `badge` (15,95%).
- **Duplicidades:** não foram encontradas duplicatas exatas nas quatro tabelas.
- **Outliers (IQR):** `Order.totalAmount` com **69 outliers**; limite inferior negativo (incoerente para valor monetário).
- **Tipagem/consistência:** campos de data/tempo (`scheduledAt`, `sendAt`) e categorias (status/IDs) exigem padronização.
- **Relacionamentos:** checar integridade entre `CampaignQueue.campaignId` ↔ `Campaign.id` e (se existir) chave **Order** ↔ **Customer**.

---

2) Melhorias prioritárias (impacto direto no Dashboard)

2.1 Missing

- **Tratar como categoria e não "chutar" valores** quando o vazio for esperado pelo processo.
    - `Order.scheduledAt`: criar categoria **"Não agendado"**.
    - `CampaignQueue.response`: criar categoria **"Sem retorno"**.

    - `CampaignQueue.sendAt`: **“Não agendado”**.
    - `Customer.enrichedAt/enrichedBy`: criar **flag Enriquecido (Sim/Não)**.
    - `Customer.gender`: manter **“Não informado”** (não imputar).
    - `Campaign.description/badge` e `Order.extraInfo`: manter **opcionais**.
- **Efeito no Dashboard:** evita perda de linhas em filtros/joins e deixa claro o motivo do vazio.

2.2 Outliers (valor monetário)

- **Order.totalAmount**
    - **Regra de negócio**: valores **≥ 0**; se negativo representar **estorno/ajuste**, separar em **coluna própria** ou **flag**.
    - **Tratamento estatístico** (se necessário): *capping* por percentis (ex.: P1–P99) apenas para visualizações agregadas.
- **Efeito no Dashboard:** KPIs de receita/ticket médio **confiáveis** e comparáveis.

2.3 Padronização de categóricas

- Normalizar caixa (tudo maiúsculo/minúsculo), remover espaços/acentos inconsistentes.
- Criar **tabelas de referência** (dicionários) p/ `status`, `channel`, `storeId`, etc.
- **Efeito no Dashboard:** filtros funcionam sem categorias duplicadas “quase iguais”.

2.4 Datas e tempos

- Validar e padronizar formatos de `scheduledAt` e `sendAt` (timezone/ISO).
- Regras simples: sem datas “impossíveis” (muito antigas/futuras), coerência ordem dos eventos.
- **Efeito no Dashboard:** séries temporais corretas e comparáveis.

2.5 Integridade de chaves (joins)

- Garantir que **`CampaignQueue.campaignId` exista em `Campaign.id`** (sem órfãos).
- Se houver chave de cliente em `Order`, validar existência em `Customer`.
- **Efeito no Dashboard:** gráficos combinados (envios por template/segmento, pedidos por perfil) sem quebras.

---

3) Plano de tratamento por tabela

Order

- **Missing:** `scheduledAt` → **"Não agendado"**; `extraInfo` → manter opcional.
- **Outliers:** regra **piso ≥ 0** em `totalAmount`; separar **estornos** (quando aplicável).
- **Datas:** padronizar `scheduledAt`; criar **derivadas** (dia da semana, hora) para análise.
- **Entrega:** dicionário atualizado + regra documentada para receita.

Customer

- **Missing:** `gender` → "Não informado"; `externalCode/enrichedAt/enrichedBy` → **flag Enriquecido**.
- **Categóricas:** padronizar `status`, `gender`.
- **Privacidade/consistência:** validar formato de `taxId`.
- **Entrega:** relatório de cobertura (% enriquecidos, % perfil completo).

CampaignQueue

- **Missing:** `response` → "Sem retorno"; `sendAt` → "Não agendado".
- **Correlação estrutural:** `id ~ jobId` quase 1 → usar **apenas um** em gráficos (evitar duplicidade).

- **Datas:** padronizar `sendAt`; derivar hora/dia para heatmaps.
- **Entrega:** taxa de resposta calculada só onde há retorno; deixar % “sem retorno” visível.

Campaign

- **Missing:** `description`, `badge` → opcionais; usar marcador “—” quando vazio no dashboard.
- **Chaves:** garantir **`id`** íntegro; revisar `templateId`/`segmentId`.
- **Entrega:** catálogo de campanhas (name, template, segmento) limpo.

---

4) Métricas de qualidade para monitorar (entrar em um “painel de dados”)

- **% Missing por coluna** (top 10) — meta: **queda** mês a mês.
- **# Registros órfãos em joins** (ex.: `campaignId` sem `Campaign`) — meta: **0**.
- **% Outliers em `totalAmount`** — meta: reduzir após regra de negócio.
- **# Categorias “quase duplicadas”** (ex.: “Ativo” x “ATIVO”) — meta: **0**.
- **Conformidade de data** (formato e faixa válida) — meta: **100%**.

---

5) Riscos se não tratar

- **KPIs distorcidos** (receita/ticket por outliers).
- **Quebra de filtros** (categorias duplicadas por caixa/acentos).
- **Perda de amostra** (joins que descartam linhas por missing).
- **Interpretações erradas** (datas incoerentes).

---

6) Roadmap sugerido (curto prazo → médio)

1. **Semana 1–2:** regras de `totalAmount` (piso/estorno) + categorias para missing (Não agendado, Sem retorno, Não informado).

2. **Semana 3–4:** padronização de categóricas + dicionário de referência (status/canal/loja).

3. **Mês 2:** integridade de chaves e validação de datas; criação das colunas derivadas de tempo.

4. **Mês 3:** painel de **métricas de qualidade** + rotina de validação automática (antes de atualizar o BI).

---

7) Conclusão

A base tem estrutura **consistente** (sem duplicatas) e lacunas **concentradas** em campos opcionais ou de integração. O principal ajuste é **formalizar regras de negócio para `totalAmount`** e **tratar missing como categorias explícitas**, garantindo **confiabilidade** nos KPIs e **usabilidade** no dashboard. Com as padronizações e checagens propostas, a qualidade evolui de forma controlada e mensurável.