Relatório da Análise de Dados da Empresa Gigamart

1. Objetivo

Este relatório tem como objetivo avaliar a base de dados de clientes da **Gigamart** segundo as principais **dimensões de qualidade dos dados** (Precisão, Atualidade, Completude, Validade, Padronização/Consistência e Unicidade).
Foram mapeados os problemas concretos e aplicadas soluções eficazes para melhorar a confiabilidade da base e possibilitar análises mais seguras.

---

2. Análise por Dimensão de Qualidade

2.1 Completude

Problema identificado:

- A coluna **CPF**, que deveria identificar cada cliente de forma única, apresentou **202 valores ausentes**.
- A coluna **Telefone** apresentou **296 valores ausentes**.

Impacto: A ausência de dados compromete a integridade da base, dificultando contato com clientes e inviabilizando análises que exigem identificadores únicos.

Solução aplicada:

- Substituímos valores ausentes por `NaN`.
- Criamos colunas auxiliares de validação (**cpf_valido** e **telefone_valido**) para sinalizar registros válidos e inválidos.

2.2 Padronização e Consistência

a) CPF

- Problema: Formatos diferentes (apenas números, números com “-” e “.”, ou incompletos).
- Solução: Padronizamos todos os CPFs como sequência de **11 dígitos numéricos**.

- Resultado: Após a padronização, foram identificados **1763 CPFs inválidos**, sendo **1561 incompletos**.
- Observação: A correção dos CPFs ausentes ou inválidos requer contato com os clientes ou integração com bases externas.

b) Estado

- Problema: Registros misturavam siglas (“PE”) e nomes por extenso (“Pernambuco”).
- Solução: Criamos um **dicionário de conversão** no Python para padronizar todas as ocorrências no formato **sigla (UF)**.
- Estabelecemos uma padronização para a adição futura de estados apenas através de sua sigla.

c) Datas

- Problema: Diferença de formatação entre **Data de Nascimento** e **Última Compra**.
- Solução: Todas as datas foram convertidas para o tipo `datetime`, a **Data de Nascimento** seguem o formato brasileiro, enquanto a **Última compra** segue o formato **AAAA-MM-DD**. Datas inválidas foram mantidas, porém validadas nas colunas de validação (**data_nascimento_valida** e **ultima_compra_valida**).
- Observação: No Google Sheets, é possível formatar para o padrão **dd/mm/aaaa** para visualização.

d) Valor Compra

- Problema: Valores estavam em formatos inconsistentes (alguns como número, outros como data).
- Solução: Forçamos a conversão de toda a coluna para valores **numéricos**, permitindo cálculos futuros. A formatação para **moeda** pode ser feita no Sheets.

e) Nome Completo

- Problema: Diferença de digitação (maiúsculas, minúsculas, títulos como “Dr.”, “Sra.”).
- Solução:
    - Padronizamos todos os nomes em minúsculas.
    - Removemos títulos extras.
    - Convertidos novamente para **formato clássico**, com a primeira letra de cada nome em maiúscula.

f) Telefone

- Problema: Formatos diferentes (com ou sem +55, DDD em formatos variados, números de SAC iniciados por 0800).
- Solução:
    - Padronizamos todos os telefones para **(DD) número**.
    - Removemos o código internacional **+55**.
    - Sinalizamos números inválidos (como os iniciados em 0800 ou incompletos) na coluna **Telefone.**
    - Adicionamos a coluna **telefone_valido** para identificar quais números são válidos e quais não são.

g) E-mail

- Problema: Presença de e-mails incompletos ou sem domínio válido.
- Solução: Criada a coluna **email_valido** para identificar registros corretos e incorretos, facilitando futuras ações de correção.

2.3 Validade

- Vários dados não estavam no **formato esperado**:
    - CPFs com menos de 11 dígitos.
    - E-mails sem domínio.
    - Telefones fora do padrão nacional.
    - Datas inválidas.
- Com as padronizações descritas, foi possível **identificar e sinalizar registros inválidos**, embora alguns ainda dependam de coleta externa para correção.

2.4 Unicidade

Problema identificado:

- Foram encontradas **60 linhas duplicadas** no DataFrame.

Solução aplicada:

- Criada a coluna **registros_duplicados**, que marca todas as linhas repetidas (inclusive a primeira ocorrência). Na planilha elas estão marcadas em amarelo na coluna dos seus ids.

---

3. Conclusão e Recomendações

A análise evidenciou que a base da Gigamart possui problemas significativos de **completude, validade, padronização e unicidade**.
As principais ações corretivas aplicadas foram:

- Padronização de CPFs, telefones, estados, nomes e valores.
- Conversão de datas para formato consistente.
- Criação de colunas auxiliares de validação (**cpf_valido, telefone_valido, email_valido, registros_duplicados, data_nascimento_valida, ultima_compra_valida**).
- Substituição de dados ausentes por `NaN`.

Próximos passos recomendados:

1. Implementar **restrições de entrada** no sistema para evitar inconsistências (ex: máscaras para CPF e telefone).
2. Criar **processos automáticos de validação contínua** (pipelines de qualidade de dados).
3. Realizar **contato com clientes** para atualizar CPFs, e-mails e telefones inválidos.
4. Analisar a coluna de **status de cliente**, identificar quais regras configuram um cliente inativo/bloqueado e analisar a necessidade de manter esses dados presentes no relatório.