

## Relatório da Verificação de Qualidade dos Dados

Neste arquivo, realizamos uma série de verificações para entender melhor a qualidade dos dados presentes nos quatro arquivos CSV

fornecidos: CampaignQueue\_semicolon.csv, Campaign\_semicolon.csv, Customer\_semicolon.csv, e Order\_semicolon.csv.

Começamos carregando cada arquivo em um DataFrame pandas (df1, df2, df3, e df4, respectivamente), especificando que o separador é o ponto e vírgula (;).

Para cada DataFrame, foram realizadas as seguintes verificações:

1. **Valores Nulos:** Verificamos a quantidade de valores nulos em cada coluna para identificar dados faltantes.
2. **Valores Únicos:** Contamos o número de valores únicos por coluna para ter uma ideia da cardinalidade dos dados.
3. **Tipos de Dados:** Verificamos os tipos de dados de cada coluna para garantir que estão no formato esperado (por exemplo, números para colunas numéricas, objetos para texto, etc.).

Além dessas verificações gerais, realizamos análises mais aprofundadas em colunas específicas de cada DataFrame:

- **df1 (FILA DE CAMPANHA):** Verificamos o formato das colunas de data/hora (createdAt, scheduledAt, sendAt, updatedAt) para garantir uniformidade no formato %d/%m/%Y %H:%M.
- **df2 (CAMPANHAS):** Verificamos o formato das colunas de data/hora (createdAt, updatedAt), esperando o formato %Y-%m-%d %H:%M:%S.
- **df3 (CLIENTES):**
  - Analisamos os valores presentes na coluna gender para entender a distribuição e identificar possíveis inconsistências, incluindo valores nulos.
  - Fizemos uma verificação básica no formato da coluna phone para identificar caracteres inesperados.
  - Verificamos a unicidade da coluna taxId (CPF/CNPJ) e utilizamos uma expressão regular básica para validar se os formatos correspondem aos esperados para CPF ou CNPJ.

- Verificamos o formato das colunas de data (dateOfBirth, enrichedAt, createdAt, updatedAt), esperando o formato %d/%m/%Y ou %d/%m/%Y %H:%M.
- **df4 (PEDIDOS):**
  - Analisamos os valores presentes na coluna status para entender os diferentes status de pedidos e a frequência de cada um.
  - Verificamos se as colunas numéricas de valores (totalAmount, preparationTime, takeOutTimeInSeconds) contêm apenas valores não negativos, assumindo que valores negativos seriam inconsistentes.
  - Verificamos a unicidade da coluna displayId para garantir que cada pedido tem um identificador único.
  - Verificamos o formato das colunas de data/hora (createdAt, scheduledAt, updatedAt), esperando o formato %d/%m/%Y %H:%M.

Em resumo, essas verificações nos permitiram identificar a presença de valores nulos, a diversidade de valores únicos, os tipos de dados e possíveis inconsistências ou formatos inesperados em colunas importantes, especialmente as de data/hora e identificadores únicos, preparando o terreno para etapas futuras de limpeza e análise dos dados.