Relatório — Exploração dos Dados

1) Objetivo da exploração  
Entender padrões, relações e distribuição das variáveis para definir métricas, filtros e visualizações do dashboard. A exploração guia:  
 - Quais KPIs mostrar (ex.: volume de pedidos, valor total, taxa de resposta de campanhas).  
- Quais dimensões de análise (ex.: por loja, por campanha, por status, por horário).  
- Que gráficos fazem sentido (histogramas, barras, heatmaps de correlação).  
- Que tratamentos precisam existir (imputação de missing, padronização de categóricas, regras para outliers).  
  
2) O que foi feito (passo a passo no Colab)  
1. Coleta: leitura dos 4 CSV (Order\_semicolon.csv, Customer\_semicolon.csv, CampaignQueue\_semicolon.csv, Campaign\_semicolon.csv) com sep=';'.  
2. Inspeção inicial: shape, head(3), tipos de dados.  
3. Exploração univariada:  
○ Numéricas: describe() + histogramas (distribuição, caudas).  
○ Categóricas: describe() + barras Top 10 (frequências).  
4. Exploração multivariada:  
○ Matriz de correlação (numéricas) para detectar colinearidade e relações fortes.  
5. Qualidade (apoio à exploração):  
○ Missing por coluna (prioriza o que precisa de regra).  
○ Duplicatas (não houve).  
○ Outliers (IQR) para checar limites realistas (principalmente monetários).  
6. Anotações: registramos achados e decisões de uso para o dashboard (o que entra como KPI, o que vira filtro, o que precisa tratamento).  
Obs.: Esses passos geram as saídas e gráficos que sustentam as conclusões abaixo.  
  
3) Principais achados por tabela e como entram no dashboard  
A) Order (pedido)  
O que vimos  
 - Distribuições: takeOutTimeInSeconds com cauda longa (grande variação de tempo de retirada).  
- Correlação: sem pares fortes entre numéricas (|corr| ≥ 0,70).  
- Qualidade: extraInfo (59,05% missing) e scheduledAt (48,40% missing). Outliers em totalAmount (69 casos; limite inferior IQR negativo).  
Como ajuda o dashboard  
- KPIs:  
○ Nº de pedidos; receita total (soma de totalAmount); ticket médio; tempo médio de retirada.  
- Dimensões/Filtros: por storeId/loja, por dia/horário (de scheduledAt quando houver), por status (se existir).  
- Gráficos:  
○ Linha/área de pedidos por dia; barras por loja/status; boxplot de tempo de retirada; histograma de totalAmount.  
- Decisões de preparo:  
○ Tratar totalAmount com piso ≥ 0 (ou separar estorno), para não distorcer KPIs.  
○ scheduledAt ausente = classificar como “não agendado” (categoria no filtro).  
  
B) Customer (cliente)  
O que vimos  
- Variabilidade: id com alta variação; status com baixa (poucas classes).  
- Missing: externalCode (62,60%), enrichedAt/enrichedBy (49,60%), gender (25,90%).  
- Correlação: nenhuma forte entre numéricas.  
Como ajuda o dashboard  
- KPIs:  
○ Nº de clientes; % de clientes enriquecidos (tem enrichedAt/enrichedBy); % de perfil completo (ex.: tem gender/taxId).  
- Dimensões/Filtros: por status, gender, flag enriquecido (sim/não).  
- Gráficos:  
○ Rosca/barras de status e gênero; indicador de cobertura de enriquecimento.  
- Decisões de preparo:  
○ Criar flag binária “enriquecido” (existe enrichedAt?) em vez de imputar.  
○ Não imputar gênero arbitrariamente; tratar como “não informado”.  
  
C) CampaignQueue (fila/envios)  
O que vimos  
- Variabilidade: id, jobId e campaignId altos.  
- Correlação: forte entre id e jobId (≈1,00).  
- Missing: response (68,24%) e sendAt (34,24%).  
Como ajuda o dashboard  
- KPIs:  
○ Total de envios; taxa de resposta (quando houver response); volume de envios por horário/dia (sendAt).  
- Dimensões/Filtros: por campanha (campaignId), loja (storeId), canal (se houver), lote (jobId).  
- Gráficos:  
○ Série temporal de envios por dia/hora; barras por campanha/loja/canal; taxa de resposta por segmento.  
- Decisões de preparo:  
○ response ausente = “sem retorno registrado” (categoria própria).  
○ sendAt ausente = “não agendado” (categoria).  
  
D) Campaign (campanha)  
O que vimos  
- Variabilidade: destaque em id, templateId, segmentId.  
- Correlação: forte entre id e templateId (≈1,00).  
- Missing: description (20,25%) e badge (15,95%).  
Como ajuda o dashboard  
- KPIs:  
○ Nº de campanhas ativas/cadastradas; distribuição por template e segmento.  
- Dimensões/Filtros: por templateId, segmentId, storeId.  
- Gráficos:  
○ Barras/treemap de campanhas por template/segmento; tabela de catálogo de campanhas.  
- Decisões de preparo:  
○ Campos textuais (description, badge) podem ficar opcionais; para o dashboard, usar rótulos curtos quando vazio (ex.: “—”).  
  
4) Como os passos levaram a esses resultados (rastreabilidade)  
- Histograma/barras mostraram distribuição (ex.: cauda longa em takeOutTimeInSeconds) → definimos boxplot e métricas de dispersão no dashboard.  
- Matriz de correlação revelou ausência/presença de colinearidade (ex.: id ~ jobId em CampaignQueue) → evitamos duplicar métricas no dashboard e usamos filtros mais úteis (por jobId).  
- Tabela de missing apontou campos críticos → decidimos por categorias “não informado/sem retorno” em vez de imputar, para manter a transparência dos dados.  
- Outliers (IQR) em totalAmount → criamos regra de piso ou separação de estornos, garantindo que KPI de receita não fique inflado ou negativo.  
  
5) Layout sugerido do Dashboard (exemplo prático)  
Página 1 — Pedidos (Order)  
- KPIs: Pedidos, Receita Total, Ticket Médio, Tempo Médio de Retirada.  
- Gráficos:  
○ Linha de Pedidos por Dia;  
○ Barras por Loja;  
○ Boxplot Tempo de Retirada;  
○ Histograma totalAmount.  
- Filtros: Loja, Status, Agendado/Não Agendado, Período.  
  
Página 2 — Clientes (Customer)  
- KPIs: Clientes, % Enriquecidos, % Perfil Completo.  
- Gráficos: Status (barra), Gênero (rosca), Cobertura de Enriquecimento (indicador).  
- Filtros: Status, Gênero, Enriquecido (Sim/Não).  
  
Página 3 — Campanhas & Envios (Campaign + CampaignQueue)  
- KPIs: Envios, Taxa de Resposta (quando disponível), Campanhas.  
-Gráficos:  
○ Linha Envios por Hora/Dia;  
○ Barras por Campanha/Template/Loja;  
○ Tabela Catálogo de Campanhas (name, template, segmento).  
- Filtros: Campanha, Template, Segmento, Loja, Canal, Lote (jobId).  
  
6) Decisões de preparação dos dados (para o BI)  
- Regras de missing: mapear como categorias explícitas (“não informado”, “sem retorno”, “não agendado”).  
- Outliers: aplicar piso ≥ 0 em totalAmount ou separar estornos.  
- Padronização de categóricas: normalizar caixa/acentos/espaços; dicionário de status/canal.  
- Chaves e joins (se necessário para o dashboard):  
○ CampaignQueue.campaignId ↔ Campaign.id (envios por campanha/template/segmento).  
○ Order ↔ Customer (se houver chave de cliente no pedido) para métricas por perfil.  
  
Conclusão  
A exploração definiu o que medir, como recortar e quais visualizações usar, além de estabelecer regras de preparo (missing, outliers, padronização) que deixam os KPIs confiáveis e o dashboard navegável.