

**FUNDAÇÃO ESCOLA DE COMÉRCIO ÁLVARES PENTEADO
FECAP
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS**

ALEXANDRA CHRISTINE SILVA RAIMUNDO - 24026156

CARLOS AUGUSTO SANTOS DE ALMEIDA - 20010535

HEBERT DOS REIS ESTEVES - 24026079

JOSÉ BENTO ALMEIDA GAMA - 24026127

**CANNOLI INTELLIGENCE
EXPLORAÇÃO E QUALIDADE DE DADOS**

**São Paulo - SP
2025**

SUMÁRIO

01. INTRODUÇÃO	3
02. EXPLORAÇÃO DOS DADOS	4
02.1. Clientes por status	4
02.2. Clientes por gênero.....	5
02.3. Clientes enriquecidos	5
02.4. Clientes por faixa etária	6
02.5. Campanhas por status	7
02.6. Campanhas por badge (tipo).....	8
02.7. Campanhas criadas por mês	9
02.8. Mensagens por status (fila)	10
02.9. Pedidos por status	11
02.10. Pedidos por canal	12
02.11. Pedidos por mês	13
02.12. Receita por mês	14
02.13. Ticket médio por canal	15
03. VERIFICAÇÃO DA QUALIDADE DOS DADOS.....	16
03.1. Valores nulos – contagem	16
03.2. Valores nulos – percentual	18
03.3. Registros duplicados	19
03.4. Valores inconsistentes em categorias	20
03.5. Integridade referencial entre tabelas	21
03.6. Unicidade de IDs (chaves primárias)	22
04. CONCLUSÃO.....	23

01. INTRODUÇÃO

Este relatório reúne a etapa de exploração e a verificação da qualidade dos dados do projeto Cannoli Intelligence. O objetivo foi entender melhor os padrões presentes nas bases Customer, Campaign, CampaignQueue e Order, além de identificar pontos que podem atrapalhar análises futuras.

A ideia foi trabalhar exploração e qualidade juntas, de forma que fosse possível ver ao mesmo tempo os insights que os dados trazem e também os problemas que precisam ser corrigidos. Para cada indicador, foi explicado o motivo da análise, os passos feitos no Google Colab (com Python, pandas e matplotlib), o resultado obtido, as limitações encontradas e os próximos passos que poderiam ser seguidos.

Os gráficos e tabelas apresentados foram retirados diretamente das execuções do notebook, o que deixa o relatório fiel ao que foi realizado. Dessa forma, o documento mostra tanto os aprendizados iniciais sobre os dados quanto os pontos que precisam de melhoria para apoiar análises futuras e a construção de dashboards.

02. EXPLORAÇÃO DOS DADOS

02.1. Clientes por status

Foi analisada a coluna status da tabela Customer com o objetivo de identificar a proporção de clientes ativos e inativos. Esse indicador é útil para entender se a base está concentrada em perfis que ainda participam das campanhas ou se há grande volume de registros antigos. Para construir o gráfico, foi utilizada a coluna status e aplicada a função `value_counts()` para obter a contagem por categoria. Em seguida, a distribuição foi exibida em gráfico de barras, por facilitar a comparação direta. O resultado evidenciou equilíbrio entre os dois grupos, sugerindo que parte da base não tem participação recente. A limitação observada é o uso de códigos numéricos (1 = Ativo; 2 = Inativo), que reduz a leitura imediata. Como próximo passo, seria adequado mapear os códigos para rótulos textuais e avaliar a possibilidade de reativação de clientes inativos por meio de campanhas específicas.

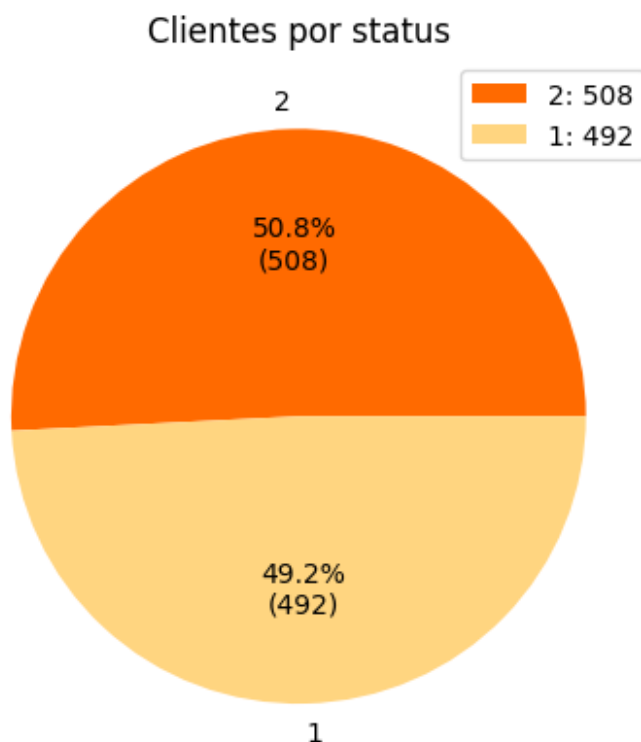


Figura 01 – Distribuição de clientes por status

02.2. Clientes por gênero

Foi analisada a coluna gender da tabela Customer para compreender o perfil demográfico e verificar inconsistências de preenchimento. A contagem foi realizada com `value_counts(dropna=False)`, incluindo valores nulos, pois a ausência de informação também carrega sinal de qualidade. Os resultados foram apresentados em gráfico de barras, por permitir comparação clara entre categorias. O gráfico mostrou predominância de 'M' e 'F', mas também a presença de 'O' (outros) e um volume relevante de NaN, o que evidencia falta de padronização. A limitação é que tais inconsistências tornam frágeis as segmentações por gênero. Como próximo passo, seria importante padronizar as categorias válidas ('M', 'F' e 'O') e considerar estratégias de enriquecimento, como inferência de gênero a partir do primeiro nome ou a manutenção explícita da categoria 'Não informado' para casos realmente sem identificação.

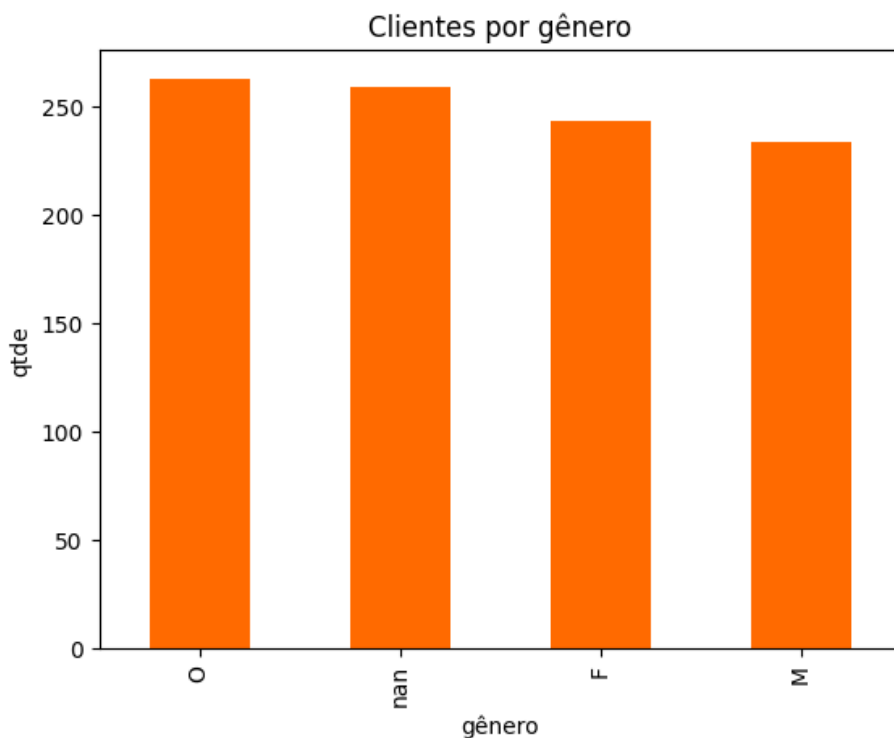


Figura 02 – Distribuição de clientes por gênero

02.3. Clientes enriquecidos

Foi analisada a coluna `isEnriched`, que sinaliza se o cadastro do cliente contém informações consideradas essenciais. O indicador foi calculado por contagem simples e exibido em gráfico de pizza para destacar a participação relativa de

cada grupo. O resultado indicou que aproximadamente metade da base está enriquecida e metade não, apontando que há espaço para completar dados faltantes antes de segmentações mais finas. A limitação é que a baixa completude restringe análises demográficas e comportamentais. Como próximo passo, seria prioritário avançar em processos de enriquecimento para idade, gênero e formas de contato, a fim de elevar a qualidade das próximas análises.

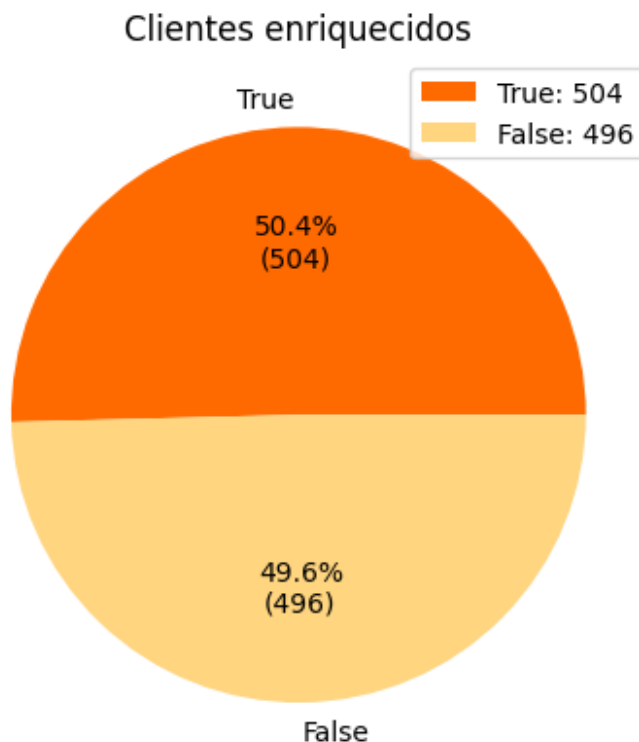


Figura 03 – Taxa de clientes enriquecidos

02.4. Clientes por faixa etária

Foi calculada a idade a partir de `dateOfBirth`, com conversão para `datetime` e subtração da data atual (anos aproximados). Em seguida, as idades foram agrupadas em faixas (0–17, 18–24, 25–34, 35–44, 45–59, 60+) via `pd.cut`, e a distribuição foi apresentada em gráfico de barras. O resultado revelou maior concentração nas faixas acima de 45 anos, com destaque para 60+, indicando um público relativamente mais maduro. A limitação é que datas inválidas ou ausentes reduzem o tamanho efetivo da amostra. Como próximo passo, seria importante revisar outliers e completar registros de nascimento para consolidar a análise etária.

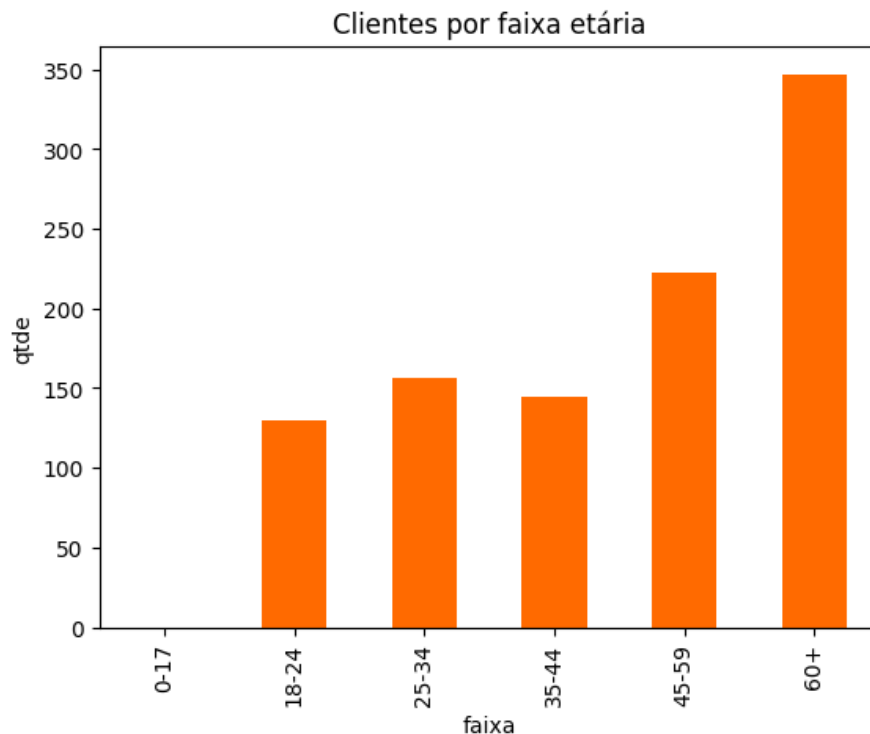


Figura 04 - Distribuição de clientes por faixa etária

02.5. Campanhas por status

Foi analisada a coluna status da tabela Campaign com a finalidade de entender a distribuição das campanhas ao longo do ciclo de vida. A contagem foi obtida com `value_counts()` e exibida em barras horizontais para facilitar a leitura dos rótulos. O resultado mostrou uma distribuição relativamente uniforme entre categorias, sem concentração extrema em um único status. A limitação, novamente, é a codificação numérica, que exige dicionário para leitura. Como próximo passo, seria interessante mapear os códigos para rótulos ('Ativa', 'Concluída', 'Cancelada') e relacionar status com métricas de resultado.

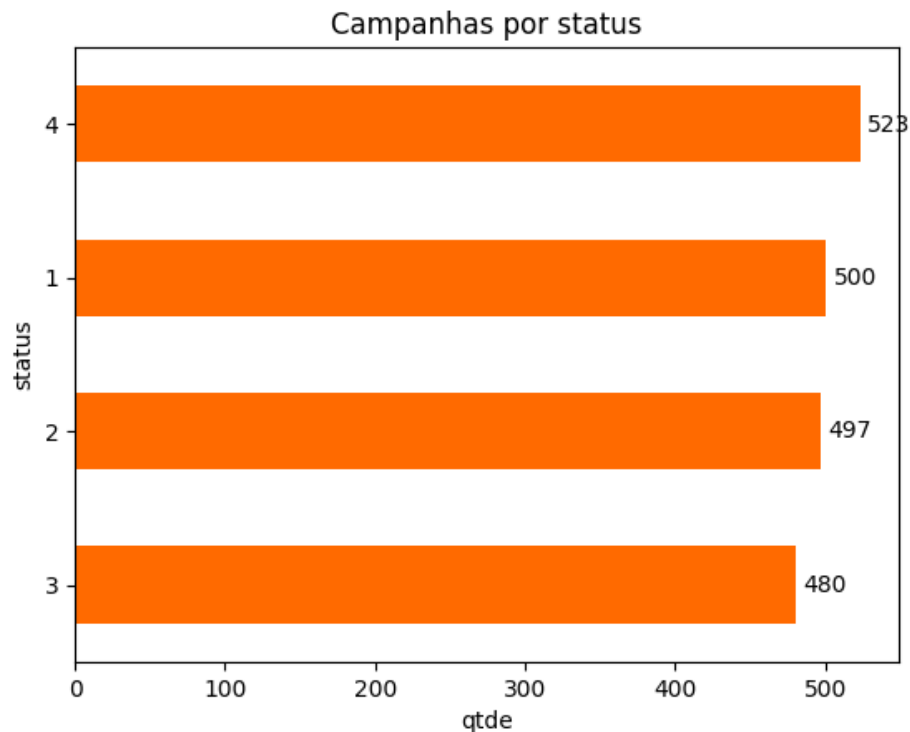


Figura 05 – Distribuição de campanhas por status

02.6. Campanhas por badge (tipo)

Foi verificada a coluna badge (tipo de campanha), substituindo previamente valores nulos por 'Não informado' para não perder a informação de lacuna. Após a contagem, os resultados foram apresentados em barras. Observou-se predominância de tipos como 'consumption' e 'winback', além de parcela relevante sem badge informado. A limitação é que a ausência do tipo compromete análises comparativas entre campanhas. Como próximo passo, seria adequado padronizar o preenchimento de badge e, em seguida, relacionar o tipo de campanha com métricas de desempenho (alcance, leitura, conversão).

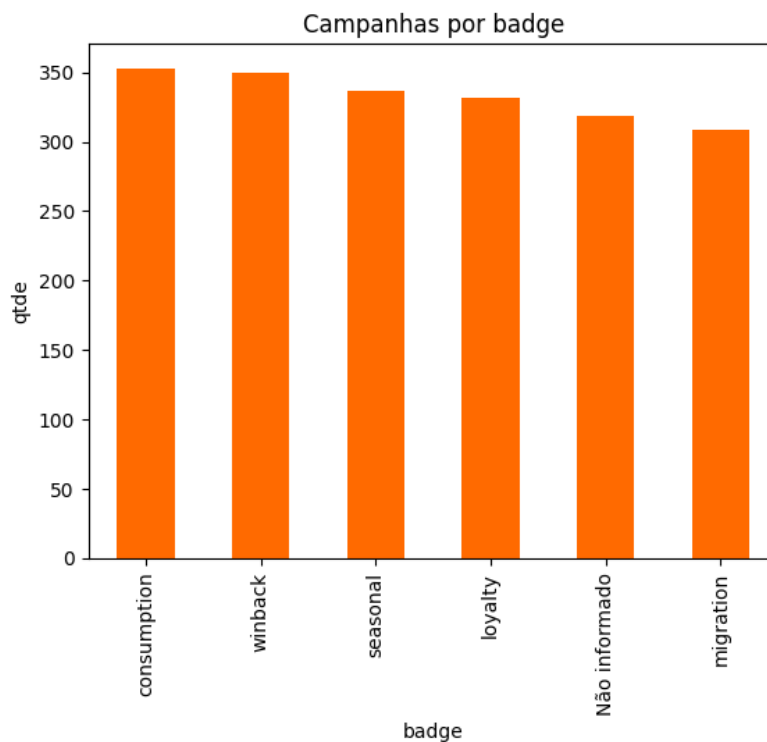


Figura 06 – Distribuição de campanhas por badge

02.7. Campanhas criadas por mês

Foi utilizada a coluna `createdAt` da `Campaign`, convertida para `datetime` e reamostrada por mês por meio de `resample('ME')`. A série mensal foi plotada em linha, com o objetivo de identificar sazonalidade de criação. O gráfico mostrou oscilações ao longo do período, com picos em meses específicos, o que pode refletir estratégias sazonais. A limitação é que meses finais podem estar parciais. Como próximo passo, seria comparar a curva de criação de campanhas com indicadores de pedidos e receita para avaliar impacto.

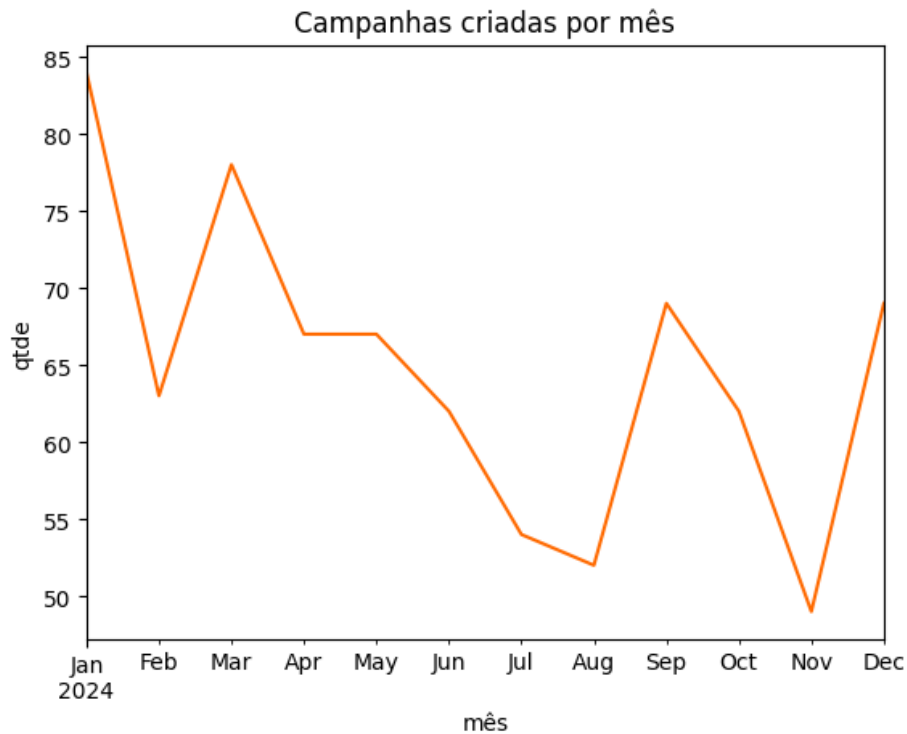


Figura 07 – Evolução mensal de campanhas criadas

02.8. Mensagens por status (fila)

Foi verificada a coluna status da CampaignQueue para compreender o comportamento da fila de mensagens. A contagem por categoria foi exibida em barras, permitindo observar a distribuição entre estados de processamento. O resultado indicou distribuição relativamente equilibrada, sugerindo que a fila tem fluidez. A limitação é a presença de códigos sem rótulo público. Como próximo passo, seria importante mapear os códigos e cruzar com horários de envio, taxa de leitura e taxa de clique para identificar gargalos.

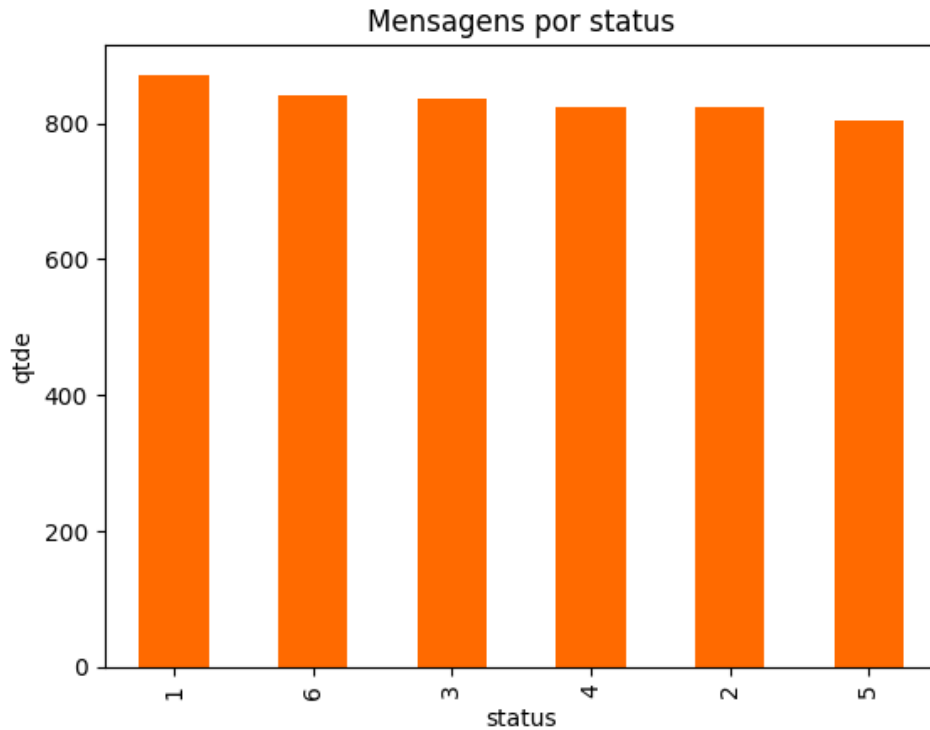


Figura 08 – Distribuição de mensagens por status

02.9. Pedidos por status

Foi analisada a coluna status da Order para entender o fluxo operacional dos pedidos. A contagem foi exibida em barras horizontais. Observou-se predominância de PENDING e DISPATCHED, com presença de CANCELLED, sinalizando potenciais perdas. A limitação é não considerar o tempo de permanência em cada status. Como próximo passo, seria calcular tempos médios e investigar motivos de cancelamento para orientar ações de melhoria.

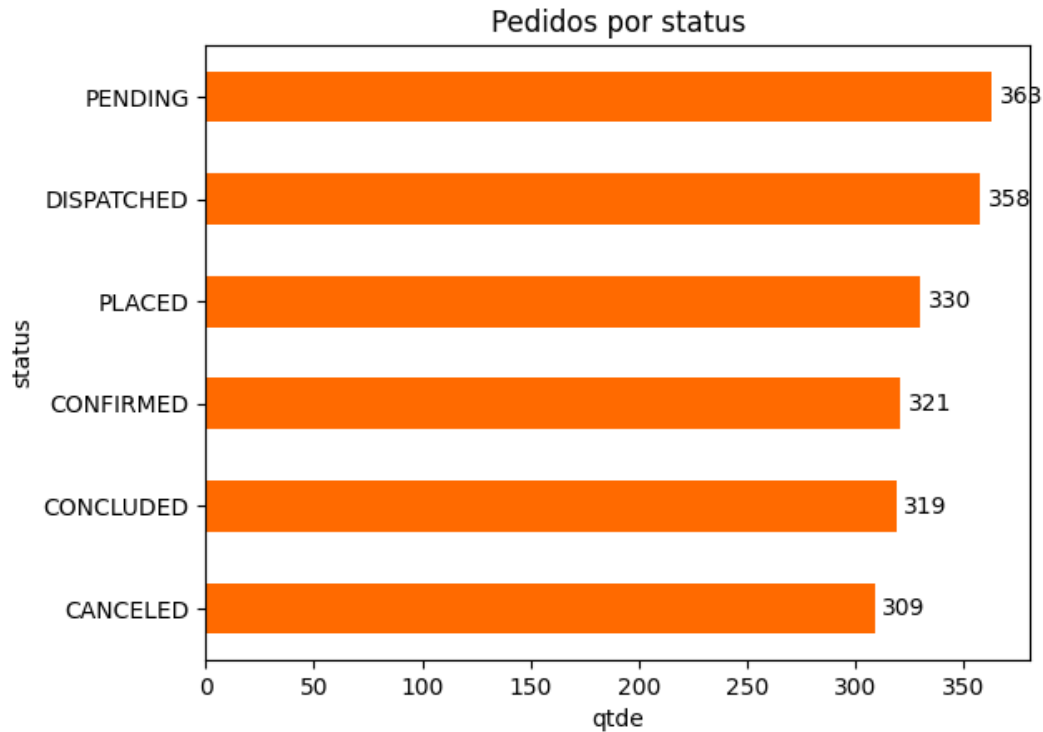


Figura 09 – Distribuição de pedidos por status

02.10. Pedidos por canal

Foi analisada a coluna salesChannel para identificar a participação relativa dos canais no volume de pedidos. A distribuição foi obtida com `value_counts()` e plotada em barras. O gráfico mostrou pedidos distribuídos por múltiplos canais, com leve predominância de EPADOCA. A limitação é que analisar apenas o volume de pedidos não traduz necessariamente a rentabilidade, já que um canal pode ter muitos pedidos de baixo valor. Como próximo passo, seria interessante cruzar os canais com métricas de ticket médio e receita para obter uma visão mais precisa sobre a rentabilidade e a relevância de cada canal.

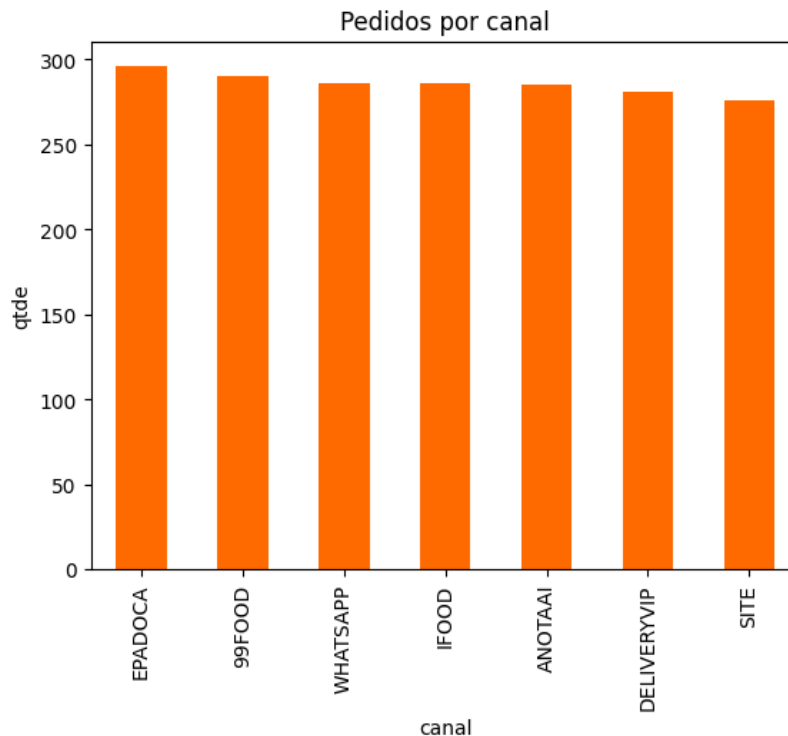


Figura 10 – Distribuição de pedidos por canal

02.11. Pedidos por mês

Foi utilizada a coluna `createdAt` da tabela `Order`, convertida para o formato `datetime` e reamostrada por mês utilizando o método `resample('ME')`. Dessa forma, obteve-se a contagem mensal de pedidos. O gráfico resultante evidenciou sazonalidade, com picos em determinados períodos do ano, indicando que a demanda não se mantém constante ao longo do tempo. Quedas acentuadas em meses finais podem estar relacionadas a janelas parciais de registro, e não necessariamente a uma redução real de pedidos. Como próximo passo, seria pertinente aplicar uma média móvel para suavizar variações bruscas e comparar a evolução mensal de pedidos com a criação de campanhas, avaliando se existe correlação entre os dois indicadores.

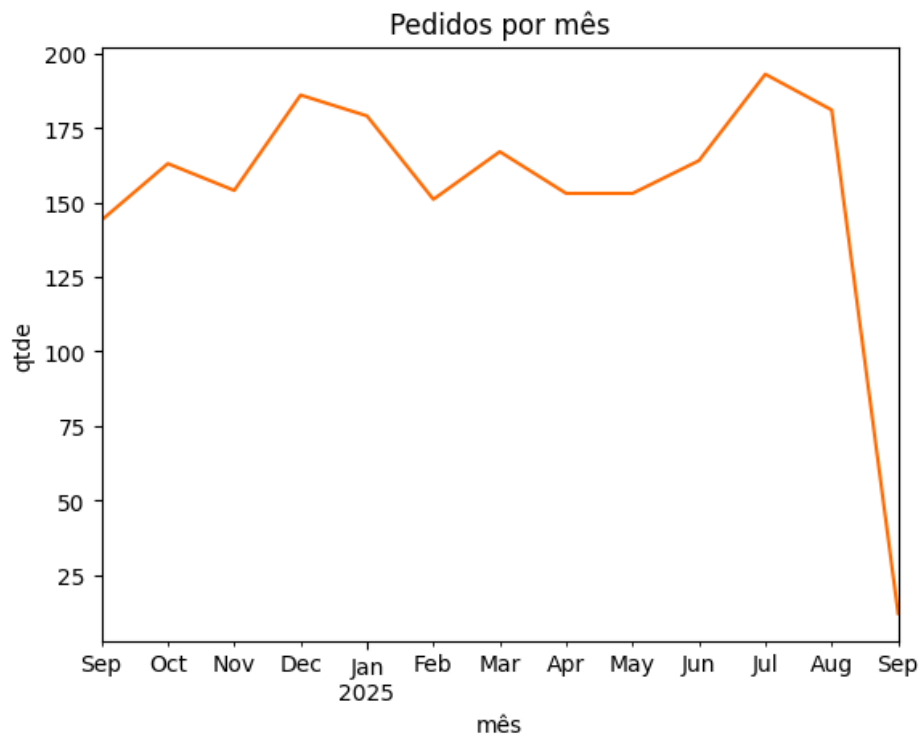


Figura 11 – Evolução mensal de pedidos

02.12. Receita por mês

Foi padronizada a coluna `totalAmount` para numérico e, na sequência, realizada a soma mensal por `resample('ME').sum()`. A série de receita acompanhou a variação de pedidos, com picos claros em alguns meses. A limitação é não distinguir receita bruta de líquida. Como próximo passo, seria separar essas métricas e avaliar o desempenho por canal para identificar fontes de maior rentabilidade.

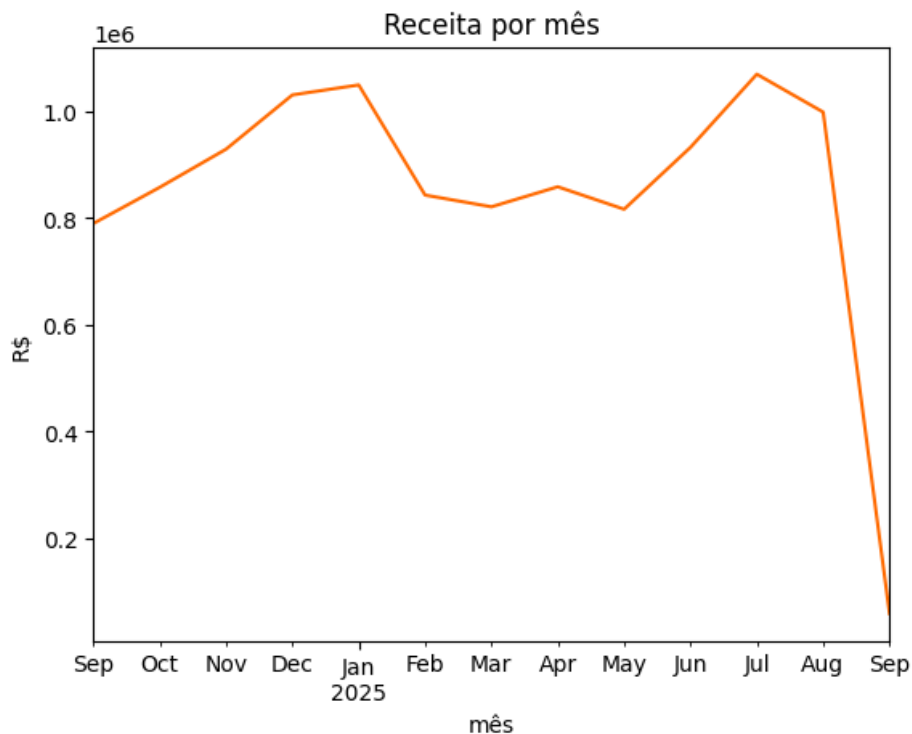


Figura 12 – Evolução mensal da receita

02.13. Ticket médio por canal

Foi calculado o ticket médio agrupando os pedidos por salesChannel e aplicando a média sobre totalAmount. A visualização em barras destacou diferenças entre canais, com ANOTAAI e iFood apresentando valores mais altos. A limitação é a sensibilidade da média a outliers. Como próximo passo, seria calcular também a mediana e o desvio padrão por canal, além de avaliar a distribuição de valores para entender a consistência do ticket.

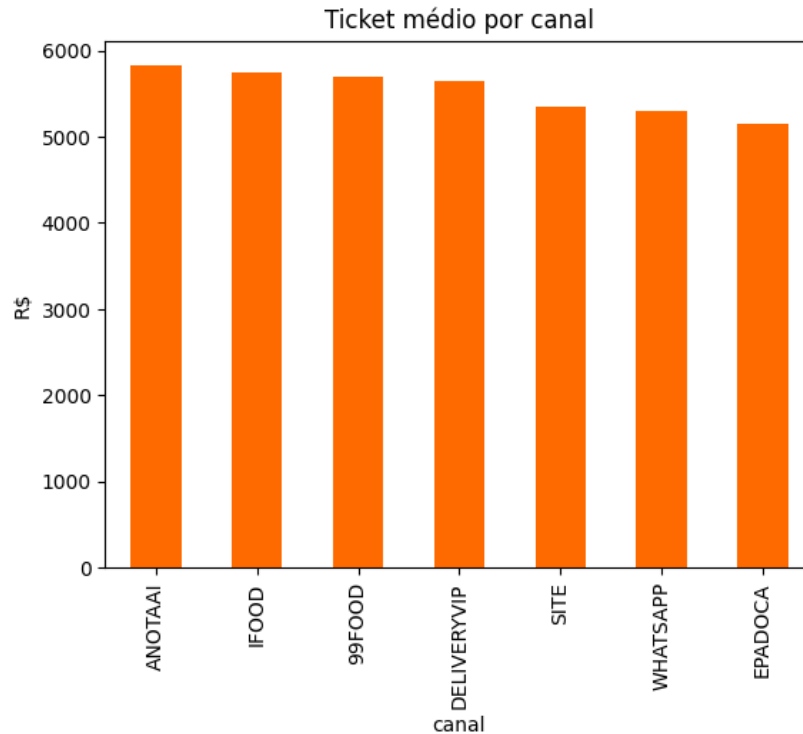


Figura 13 – Ticket médio por canal

03. VERIFICAÇÃO DA QUALIDADE DOS DADOS

03.1. Valores nulos – contagem

Foi aplicada a função `isna().sum()` em todas as tabelas para medir a quantidade absoluta de valores nulos por coluna. Os resultados (Figuras 14 e 15) mostraram lacunas relevantes. Em Order, destacaram-se `extraInfo` (1181 nulos) e `scheduledAt` (968 nulos). Em Campaign, os maiores problemas ocorreram em `description` (405), `badge` (319) e `_createdAt` (1224). Em CampaignQueue, os campos `sendAt` (1712) e `response` (3412) apresentaram ausências expressivas. Em Customer, o campo `gender` teve 259 ausentes, além de problemas em `externalCode` (626) e variáveis ligadas ao enriquecimento. Essa análise de contagem ajuda a identificar colunas críticas e orientar prioridades iniciais de tratamento.

NULOS (contagem)

Order:	
id	0
companyId	0
containerId	0
createdAt	0
customer	0
displayId	0
engineId	0
engineName	0
engineType	0
extraInfo	1181
integrated	0
integrationId	0
isTest	0
orderTiming	0
orderType	0
salesChannel	0
scheduledAt	968
status	0
preparationTime	0
takeOutTimeInSeconds	0
totalAmount	0
updatedAt	0
version	0
_createdAt	0
dtype: int64	
Campaign:	
id	0
segmentId	0
templateId	0
storeId	0
name	0
description	485
badge	319
type	0
status	0
isDefault	0
createdAt	0
createdBy	0
updatedAt	0
updatedBy	0
_createdAt	1224
dtype: int64	

Figura 14 – Contagem de valores nulos por coluna (Order e Campaign)

CampaignQueue:	
id	0
jobId	0
campaignId	0
storeId	0
storeInstanceId	0
customerId	0
phoneNumber	0
scheduledAt	0
sendAt	1712
status	0
message	0
response	3412
createdAt	0
createdBy	0
updatedAt	0
updatedBy	0
_scheduledAt	0
dtype: int64	
Customer:	
id	0
name	0
taxId	0
gender	259
dateOfBirth	0
status	0
externalCode	626
isEnriched	0
enrichedAt	496
enrichedBy	496
createdAt	0
createdBy	0
updatedAt	0
updatedBy	0
phone	0
email	0
dtype: int64	

Figura 15 – Contagem de valores nulos por coluna (CampaignQueue e Customer)

03.2. Valores nulos – percentual

Foi calculada a proporção de valores nulos em relação ao total de registros (Figuras 16 e 17). Os resultados evidenciaram que algumas colunas apresentam níveis muito altos de incompletude: em Campaign, `_createdAt` atinge 61,2% de ausências; em CampaignQueue, `response` representa 68,2% de nulos; em Customer, `externalCode` chega a 62,6%. Esses percentuais indicam que certas variáveis, se não forem tratadas, inviabilizam análises robustas. Como próximo passo, pode-se aplicar imputação quando tecnicamente viável, padronizar registros ausentes como “Não informado” em campos categóricos e acompanhar periodicamente a evolução desses percentuais como métrica de governança de dados.



Figura 16 – Percentual de valores nulos por coluna (Order e Campaign)

```

CampaignQueue (%):
  id          0.00
  jobId       0.00
  campaignId  0.00
  storeId     0.00
  storeInstanceId 0.00
  customerId  0.00
  phoneNumber 0.00
  scheduledAt 0.00
  sendAt      34.24
  status      0.00
  message     0.00
  response    68.24
  createdAt   0.00
  createdBy   0.00
  updatedAt   0.00
  updatedBy   0.00
  _scheduledAt 0.00
  dtype: float64

Customer (%):
  id          0.0
  name        0.0
  taxId       0.0
  gender      25.9
  dateOfBirth 0.0
  status      0.0
  externalCode 62.6
  isEnriched  0.0
  enrichedAt  49.6
  enrichedBy  49.6
  createdAt   0.0
  createdBy   0.0
  updatedAt   0.0
  updatedBy   0.0
  phone       0.0
  email       0.0
  dtype: float64

```

Figura 17 – Percentual de valores nulos por coluna (CampaignQueue e Customer)

03.3. Registros duplicados

Foi aplicada a função `duplicated().sum()` em todas as tabelas para verificar a existência de registros duplicados. O resultado (Figura 18) mostrou que nenhuma das tabelas apresentou duplicidades, indicando que os dados estão consistentes nesse aspecto. Esse é um ponto positivo, pois elimina o risco de inflar contagens de clientes, pedidos ou campanhas por repetição de registros.

Apesar da ausência de duplicados, a etapa é relevante para garantir confiabilidade na análise, já que bases reais frequentemente apresentam esse tipo de problema. Como próximo passo, seria pertinente manter esse monitoramento em processos futuros de ingestão de dados, aplicando validações de unicidade em campos-chave como `id`, `customerId` e `campaignId`.

```

DUPLICADOS
Order: 0
Campaign: 0
CampaignQueue: 0
Customer: 0

```

Figura 18 – Verificação de registros duplicados (Order, Campaign, CampaignQueue e Customer)

03.4. Valores inconsistentes em categorias

Foi avaliada a consistência de variáveis categóricas nas diferentes tabelas, buscando identificar se os valores presentes estão padronizados e se permitem interpretações diretas.

Na tabela Customer, a variável gender apresentou registros em três categorias principais — 'M', 'F' e 'O' (Outros) — além de valores ausentes. A presença da categoria 'O' não representa um erro, pois pode se referir a pessoas que se identificam fora do binário tradicional. No entanto, a ausência de informação em muitos registros compromete segmentações demográficas. Já a coluna status dessa tabela apresenta apenas os códigos 1 e 2, cuja interpretação não é evidente sem um dicionário de dados, dificultando análises mais rápidas.

Na tabela Order, as variáveis categóricas se mostraram mais claras. A coluna status traz valores textuais como PENDING, CONFIRMED, PLACED, DISPATCHED, CONCLUDED e CANCELED, todos em inglês e padronizados em letras maiúsculas, o que facilita a consistência. Já a variável salesChannel apresentou diferentes canais de venda, como ANOTAAL, WHATSAPP, EPADOCA, 99FOOD, IFOOD, SITE e DELIVERVIP, também padronizados em maiúsculas, o que evita problemas de variação.

Na tabela Campaign, o campo badge apresentou categorias claras como winback, loyalty, migration, seasonal e consumption. Entretanto, a coluna status utiliza apenas códigos numéricos (1, 2, 3, 4). Esse mesmo padrão é encontrado em CampaignQueue, cuja coluna status contém valores de 1 a 6. Nesses casos, a ausência de mapeamento explícito dificulta a interpretação.

De forma geral, variáveis categóricas em formato textual (como badge, gender e salesChannel) oferecem maior clareza, ainda que precisem de enriquecimento para reduzir ausências, enquanto variáveis representadas por códigos numéricos (status em Campaign, CampaignQueue e Customer) exigem documentação complementar para viabilizar análises diretas. Como próximo passo, será necessário criar um dicionário de categorias aceitas para cada campo, mapeando os códigos numéricos para rótulos textuais, tratar ausências de forma explícita utilizando categorias como Não informado e complementar os campos críticos, especialmente o de gênero, de modo a permitir análises segmentadas mais robustas e consistentes.

```

Order (categorias)
status: ['DISPATCHED' 'CONCLUDED' 'CANCELED' 'CONFIRMED' 'PENDING' 'PLACED']
salesChannel: ['ANOTAAL' 'WHATSAPP' 'EPADUCA' '99FOOD' 'SITE' 'IFOOD' 'DELIVERYVIP']

Campaign (categorias)
status: [4 3 2 1]
badge: ['winback' 'loyalty' 'migration' 'seasonal' 'consumption']

CampaignQueue (categorias)
status: [5 4 1 6 2 3]

Customer (categorias)
status: [1 2]
gender: ['O' 'M' 'F']

```

Figura 19 – Valores inconsistentes em variáveis categóricas

03.5. Integridade referencial entre tabelas

Foi verificada a integridade referencial entre tabelas para assegurar que os relacionamentos estivessem corretos. Em CampaignQueue, os campos campaignId e customerId foram comparados com as chaves primárias id das tabelas Campaign e Customer, utilizando verificações de pertinência por meio do método `isin()`. O resultado, ilustrado na Figura 20, demonstrou que não há registros sem correspondência, ou seja, todas as mensagens possuem vínculos válidos com campanhas e clientes cadastrados.

Esse resultado reforça a consistência relacional da base de dados e garante maior confiabilidade para análises que dependem de múltiplas tabelas. Apesar disso, a ausência de problemas não elimina a necessidade de prevenção futura. Como próximo passo, seria importante implementar restrições de chave estrangeira diretamente na origem, garantindo que novos registros só possam ser inseridos se houver correspondência válida. Além disso, a manutenção de processos periódicos de auditoria é recomendada para detectar eventuais falhas de integração ao longo do tempo.

```

CampaignQueue com campaignId inexistente: 0
CampaignQueue com customerId inexistente: 0


```

Figura 20 – Integridade referencial entre CampaignQueue, Campaign e Customer

03.6. Unicidade de IDs (chaves primárias)

Foi verificada a unicidade das chaves primárias (id) em todas as tabelas, com o objetivo de confirmar se cada registro possuía um identificador exclusivo. Essa checagem é fundamental para evitar ambiguidades em relacionamentos e garantir a integridade dos dados. A Figura 21 mostra que todos os IDs são únicos e não há registros duplicados, ou seja, não foram encontradas violações de unicidade em nenhuma das tabelas analisadas.

Esse resultado confirma que a base segue boas práticas de modelagem, já que a ausência de colisões nos identificadores assegura consistência nas referências cruzadas entre tabelas. Apesar disso, é recomendável manter validações automáticas de unicidade e implementar restrições de chave primária no banco de dados de origem, de forma a prevenir problemas futuros.



```
Unicidade de IDs
id: 2000 únicos de 2000 linhas (duplicados: 0)
id: 2000 únicos de 2000 linhas (duplicados: 0)
id: 5000 únicos de 5000 linhas (duplicados: 0)
id: 1000 únicos de 1000 linhas (duplicados: 0)
```

Figura 21 – Verificação de unicidade dos IDs principais

04. CONCLUSÃO

Portanto, a exploração permitiu identificar padrões de negócio relevantes, como o equilíbrio entre clientes ativos e inativos, a predominância de um público mais maduro, a sazonalidade presente em campanhas e pedidos e as diferenças de comportamento entre canais de venda. Paralelamente, ficou evidente que a confiabilidade das análises depende diretamente da qualidade do preenchimento dos dados.

A verificação de qualidade confirmou pontos frágeis, como a presença de valores nulos em campos estratégicos (por exemplo, gender e badge) e a ocorrência de categorias pouco padronizadas, que limitam interpretações mais avançadas. Por outro lado, verificou-se que os mecanismos de unicidade e a integridade referencial estão preservados, o que garante maior segurança nas relações entre tabelas e consistência dos identificadores.

Como próximos passos, seria adequada a priorização da redução de nulos, por meio de enriquecimento e padronização; a criação de dicionários de categorias para evitar divergências; e o monitoramento contínuo da completude dos dados. Essas medidas podem elevar significativamente a maturidade analítica e a confiabilidade do conjunto de dados.