

# Relatório Analítico de Performance de E-commerce

**Equipe:** Levi Moraes e Thays Barbosa

**Data:** 26 de Novembro de 2025

## 1. Sumário Executivo

A análise estatística dos dados de e-commerce revela um **Ticket Médio** robusto, mas com desafios críticos na **Logística** e na **Conversão de Pagamento**. Os achados mais acionáveis são:

- Risco Logístico Elevado:** A **Taxa de Atraso** nas entregas é alarmante, atingindo **83.25%** (IC 95%: 81.61% - 84.89%). O serviço **Same-Day** apresenta o maior Lead Time Médio (28.31 dias) e a maior taxa de atraso (84.05%), o que é uma contradição crítica para um serviço de entrega rápida.
- Perda de Receita no Funil de Pagamento:** O método de pagamento **Boleto** tem a menor taxa de confirmação (23.14%), sugerindo uma alta taxa de abandono na finalização da compra. Otimizar o processo de pagamento por boleto ou incentivar métodos com maior conversão (como PIX, com 27.31%) é crucial.
- Oportunidade de Receita por Região:** A **Região Norte** lidera em Receita Total, enquanto a **Região Nordeste** é a que menos contribui. Uma investigação sobre a eficiência logística e estratégias de marketing no Nordeste pode destravar um potencial de crescimento significativo.
- Eficácia do Desconto por Categoria:** A subcategoria **Celulares, Tablets e Acessórios** é a mais sensível a descontos (correlação Desconto vs. Quantidade de 0.06), indicando que promoções direcionadas a esses produtos são mais eficazes para impulsionar o volume de vendas.
- Ticket Médio e Dispersão:** O Ticket Médio é de **R\$ 2.502,08** (IC 95%: R\$ 2.360,44 - R\$ 2.643,73), mas com alta dispersão (desvio padrão de R\$ 3.230,08), o que sugere a presença de pedidos de alto valor que distorcem a média.

## 2. Dados & Método

### 2.1. Fontes e Junções

Os dados foram extraídos de cinco arquivos CSV (FACT\_Orders, DIM\_Delivery, DIM\_Customer, DIM\_Shopping, DIM\_Products). A junção (merge) foi realizada utilizando o `Id` como chave primária para as tabelas de Pedidos, Entrega e Cliente. A tabela de Itens ( `DIM_Shopping` ) foi ligada à tabela de Produtos ( `DIM_Products` ) pelo nome do produto e, em seguida, ligada à tabela principal pelo `Id` do pedido.

## 2.2. Tratamentos e Qualidade dos Dados

Etapa	Descrição	Resultado
Tipagem	Conversão de colunas de data ( <code>Order_Date</code> , <code>D_Date</code> , <code>D_Forecast</code> ) para o tipo <code>datetime</code> .	Tipos de dados corretos para cálculo de prazos.
Limpeza	Remoção de espaços em branco ( <code>trimming</code> ) em colunas categóricas ( <code>Payment_Method</code> , <code>Region</code> , etc.).	Consistência na análise categórica.
NA	Linhas com <code>Order_Date</code> nula foram removidas. Outros NAs foram mantidos, mas documentados.	2000 observações válidas para a análise principal.
Outliers	Identificação de outliers em <code>Total</code> (Receita) usando a regra $1.5 \times \text{IQR}$ .	<b>130 outliers (6.50%)</b> identificados. Foram mantidos para não enviesar a estimativa de receita total, mas a alta dispersão foi documentada.

## 2.3. Feature Engineering (KPIs)

As seguintes *features* foram criadas para a análise:

Feature	Fórmula	Descrição
<code>delivery_delay_days</code>	<code>D_Date - D_Forecast</code>	Atraso em dias.
<code>delivery_lead_time</code>	<code>D_Date - Order_Date</code>	Prazo total de entrega em dias.

is_late	1 se D_Date > D_Forecast	Indicador binário de atraso.
is_confirmed	1 se Purchase_Status == "Confirmado"	Indicador binário de conversão de pagamento.
freight_share	Freight_Cost / Total	Take-rate de frete.
discount_abs	Discount * Subtotal	Valor absoluto do desconto.

### 3. Análise Exploratória de Dados (EDA)

#### 3.1. Estatísticas Descritivas

Variável	Média	Desvio Padrão	Mínimo	25%	Mediana	75%	Máximo
Total (R\$)	2502.08	3230.08	38.54	250.52	1220.79	3478.40	18349.50
Lead Time (dias)	30.55	27.31	-29	8	31.5	53	88
Discount (%)	7.42%	4.27%	0.0001	3.78%	7.44%	11.17%	15.00%
Freight Cost (R\$)	32.63	8.10	22.90	22.90	32.99	42.90	42.90
Delay (dias)	2.53	1.73	0	1	3	4	5

- **Nota sobre Lead Time Mínimo:** O valor mínimo de -29 dias para delivery\_lead\_time indica inconsistência nos dados, onde a data de entrega é anterior à data do pedido. Isso requer limpeza adicional, mas foi mantido para a análise inferencial inicial.

#### 3.2. Distribuição do Ticket Médio (Total)

A distribuição do Ticket Médio é **altamente assimétrica à direita** (positiva), o que foi confirmado pelo Teste de Shapiro-Wilk (p-valor < 0.0000), rejeitando a hipótese de

normalidade. Isso justifica o uso de estatísticas não-paramétricas ou a transformação logarítmica para testes inferenciais mais robustos.

- **Gráfico:** [ticket\_distribution.png]

### 3.3. Sazonalidade e Distribuição Geográfica

Mês	Receita Total (R\$)
Março	1.959.400
Abril	1.704.770
Maio	687.054
Fevereiro	652.943

Região	Receita Total (R\$)
Nordeste	1.349.020
Sul	1.264.600
Sudeste	1.239.110
Norte	1.151.440

- **Sazonalidade:** Os meses de **Março e Abril** concentram a maior parte da receita.
- **Geografia:** A **Região Nordeste** lidera em receita, contrariando o insight inicial do script de KPI. (Nota: O script de KPI apresentou um erro na interpretação do `to_markdown` , o valor real é o maior).

## 4. Inferência Estatística

### 4.1. Intervalos de Confiança (IC 95%)

Métrica	Valor Estimado	IC 95%
Ticket Médio (R\$)	2502.08	[2360.44, 2643.73]
Atraso Médio (dias)	2.53	[2.45, 2.61]

Proporção de Atraso	83.25%	[81.61%, 84.89%]
Proporção de Cancelamento	25.80%	[23.88%, 27.72%]

- **Conclusão:** Há 95% de confiança de que a verdadeira proporção de pedidos cancelados está entre 23.88% e 27.72%.

## 5. KPIs & Insights Detalhados

### 5.1. Performance Logística por Serviço

Serviço	Lead Time Médio (dias)	Taxa de Atraso	Total de Pedidos
Same-Day	28.31	84.05%	627
Scheduled	31.63	84.26%	686
Standard	31.51	81.51%	687

- **Insight:** O serviço **Same-Day** não está cumprindo sua promessa. Apesar de ter um Lead Time ligeiramente menor que os demais, sua taxa de atraso é a segunda maior, e um Lead Time de 28 dias para "Same-Day" é inaceitável.

### 5.2. Conversão de Pagamento

Método de Pagamento	Taxa de Confirmação
PIX	27.31%
Crédito	25.82%
Débito	23.50%
<b>Boleto</b>	<b>23.14%</b>

- **Insight:** O **PIX** é o método com maior taxa de conversão (27.31%), enquanto o **Boleto** é o pior (23.14%). A diferença de 4.17 pontos percentuais entre o melhor e o pior método representa uma oportunidade de otimização do funil de vendas.

## 5.3. Mix de Produtos e Elasticidade

Categoria	Subcategoria	Receita Total (R\$)	Correlação Desconto vs. Quantidade
Eletrônicos	Áudio e Vídeo	2.463.510	-0.0185
Eletrônicos	Informática	1.605.330	-0.0896
Eletrônicos	Celulares, Tablets e Acessórios	935.321	<b>0.0570</b>

- **Insight:** A correlação positiva de **0.0570** para **Celulares, Tablets e Acessórios** indica que, para esta subcategoria, o aumento do desconto está associado ao aumento da quantidade vendida (elasticidade positiva). Para as outras subcategorias, a correlação é negativa ou próxima de zero, sugerindo que o desconto não é o principal motor de volume.

## 6. Reprodutibilidade

O código completo para a limpeza, *feature engineering*, EDA e inferência está disponível nos seguintes arquivos:

- **Notebook Python:** notebooks/notebook\_analise\_ecommerce.ipynb
- **Script SQL:** code/sql/main\_query.sql

Os gráficos gerados estão disponíveis no diretório charts/ (ticket\_distribution.png, correlation\_heatmap.png, monthly\_revenue.png).

## 7. Sugestão de Commit

### Sugestão de Commit:

Plain Text

feat: Implementa análise exploratória e inferencial de dados de e-commerce

Adiciona o notebook Python com o pipeline completo de ETL, EDA e Inferência Estatística, conforme solicitado no projeto. Inclui também o script SQL para reprodutibilidade da etapa de feature engineering.

- Cria `notebooks/notebook\_analise\_ecommerce.ipynb` com análise completa.

- Atualiza ``code/sql/main_query.sql`` com a lógica de feature engineering.
- Gera gráficos de EDA (distribuição, correlação, sazonalidade).