

NOME DA INSTITUIÇÃO: Centro Universitario Santo Agostinho

CURSO: Engenharia de Software

DISCIPLINA: Ciência de Dados

PROFESSORA: Heloisa Guimarães Coelho

Alunos: João Pedro Lima Barbosa, Rodofo Dheymison Ferreira Silva

Turma: 28M4A

Data: 01/12/2025

Trabalho Final: Análise e Pré-Processamento de Dados com o Dataset Olist E-Commerce

1. Contextualização do Problema

O comércio eletrônico brasileiro cresce de forma acelerada, mas enfrenta desafios logísticos que impactam diretamente a experiência do cliente. Fatores como atraso na entrega, valor do frete, características do produto e tempo de processamento são determinantes para a satisfação final.

Este trabalho analisa dados reais da Olist, maior e-commerce brasileiro no modelo marketplace, com o objetivo de:

- compreender o comportamento dos pedidos, entregas e produtos,
- identificar gargalos logísticos,
- analisar padrões de preço e frete,
- investigar atrasos e volume de vendas por categoria,
- construir um dataset totalmente limpo e padronizado,
- aplicar técnicas completas de pré-processamento e feature engineering.

2. Apresentação dos Datasets Utilizados

Foram utilizados **três datasets**:

2.1 olist_orders_dataset.csv

Contém informações sobre:

- status do pedido,
- datas de compra, aprovação, envio e entrega,
- data estimada de entrega.

2.2 olist_order_items_dataset.csv

Contém:

- produto vendido,
- preço,
- valor do frete,
- quantidade de itens.

2.3 olist_products_dataset.csv

Contém:

- dimensões físicas dos produtos,
- peso,
- quantidade de fotos,
- categoria do produto.

Cada dataset foi analisado individualmente quanto a:

- linhas e colunas,
- tipos de dados,
- valores ausentes,
- consistência estrutural.

3. Aplicação do Ciclo de Vida da Ciência de Dados

O trabalho segue todas as etapas do ciclo de vida:

3.1 Entendimento do Problema

Contextualização do e-commerce e seus desafios.

3.2 Entendimento dos Dados

Carregamento dos três datasets e análise inicial (shape, tipos, head()).

3.3 Preparação dos Dados

Tratamento de:

- valores ausentes,
- inconsistências,
- outliers,
- padronização textual,
- conversão de tipos.

3.4 Exploração dos Dados

Geração de gráficos e análises descritivas.

3.5 Modelagem / Pré-processamento

Aplicação de:

- codificação,
- normalização,
- seleção de atributos,
- feature engineering,
- pipeline consolidado.

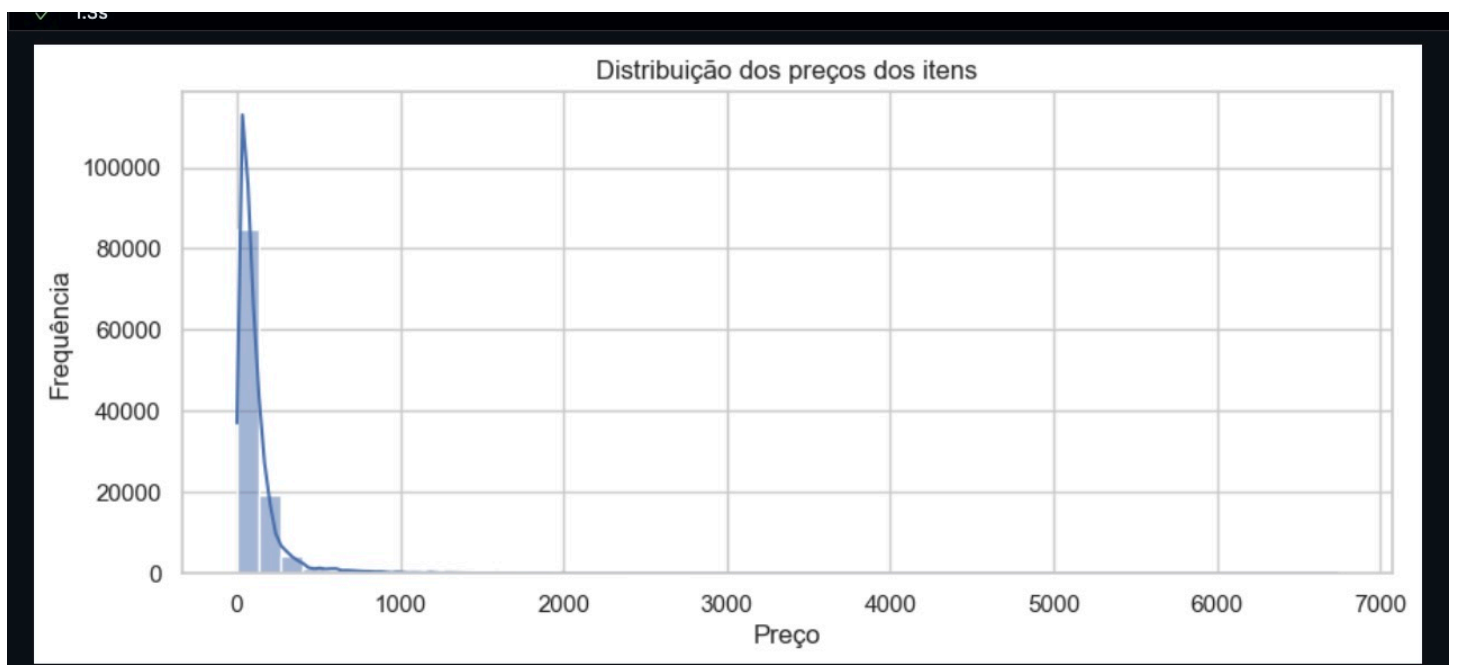
3.6 Comunicação dos Resultados

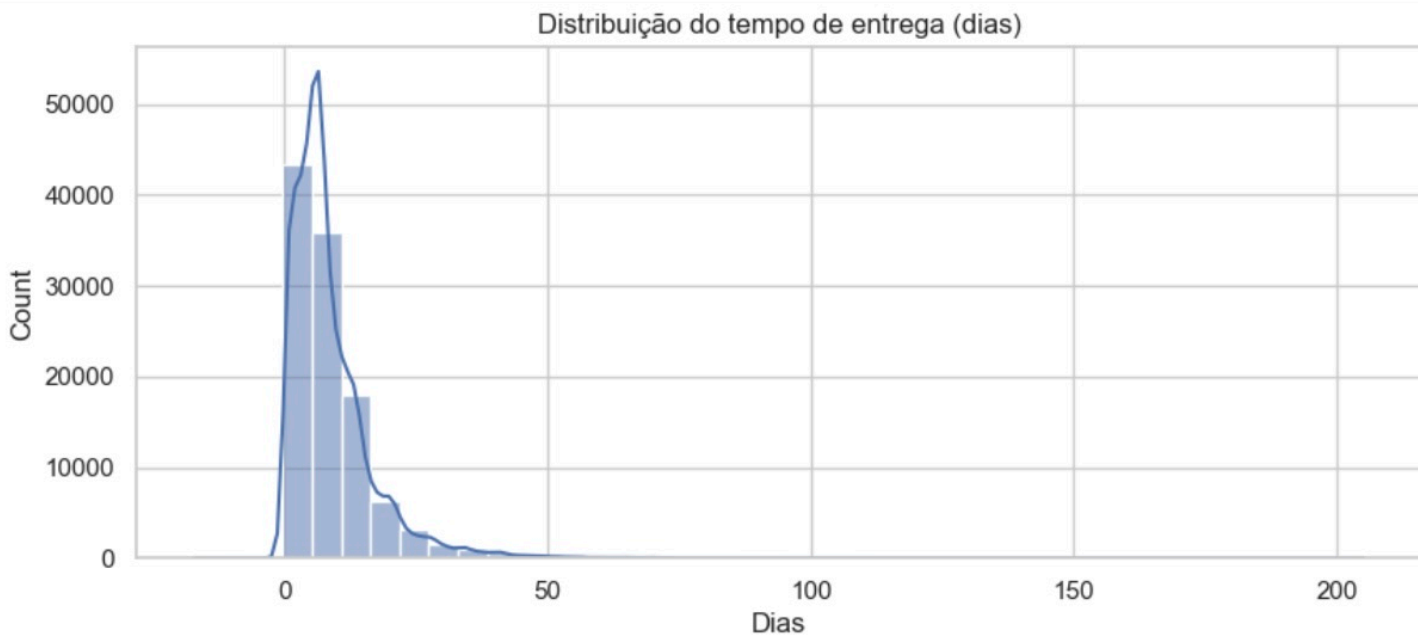
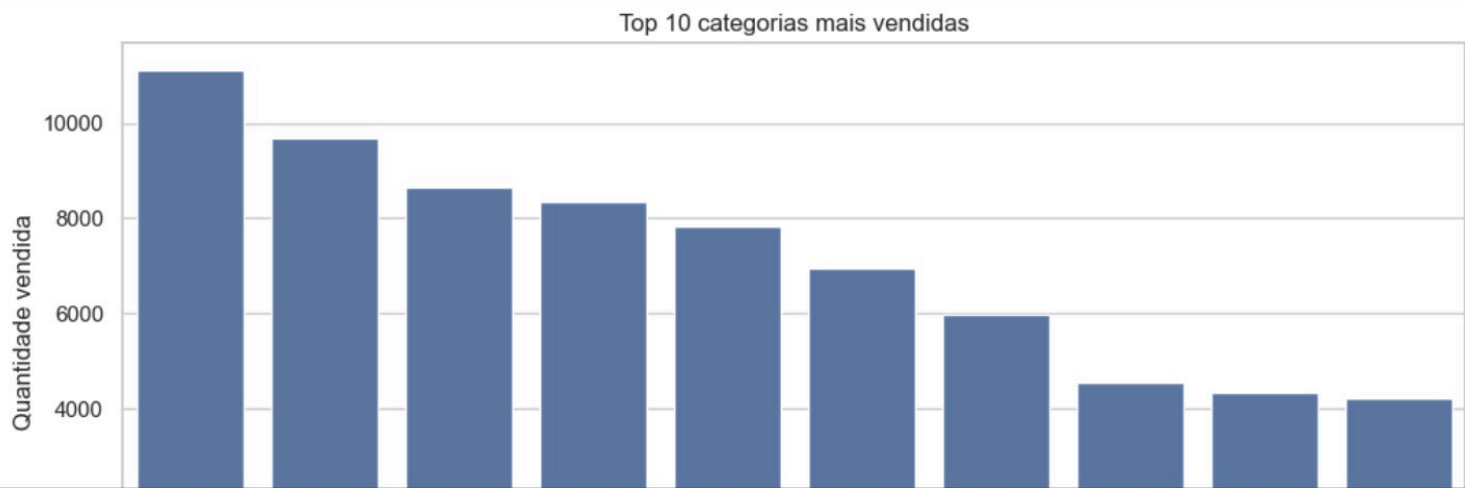
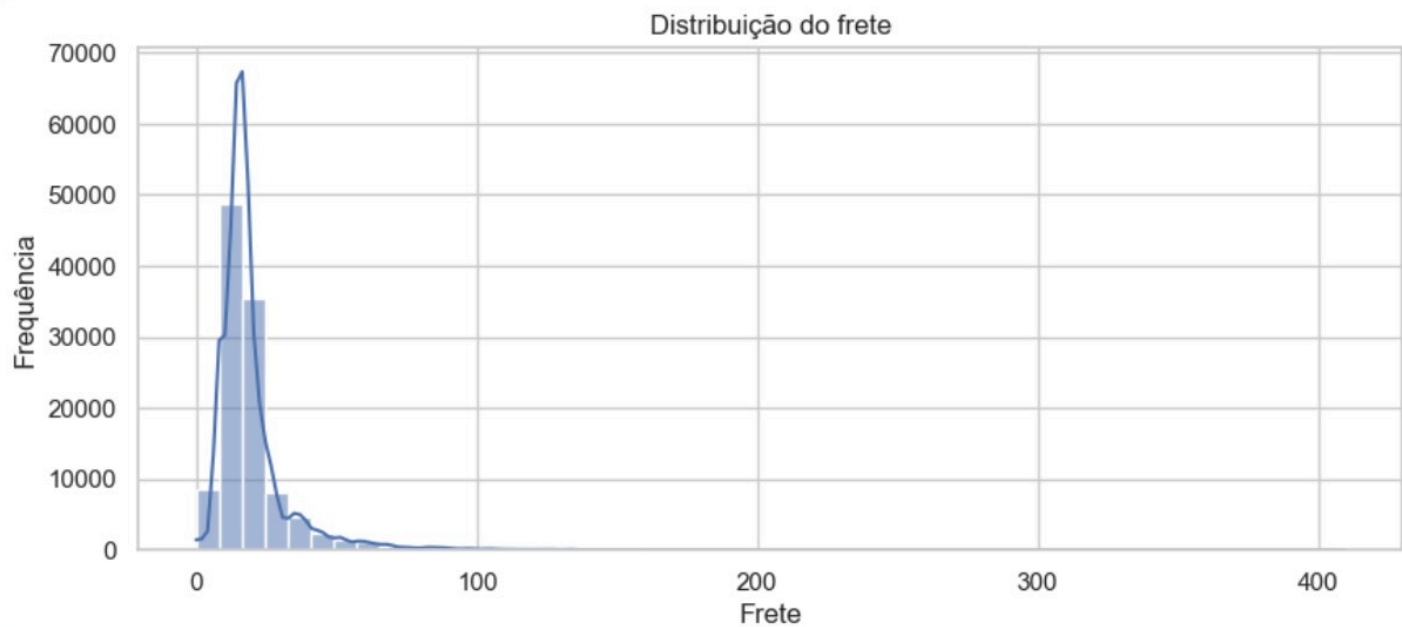
Geração de insights, gráficos e conclusões.

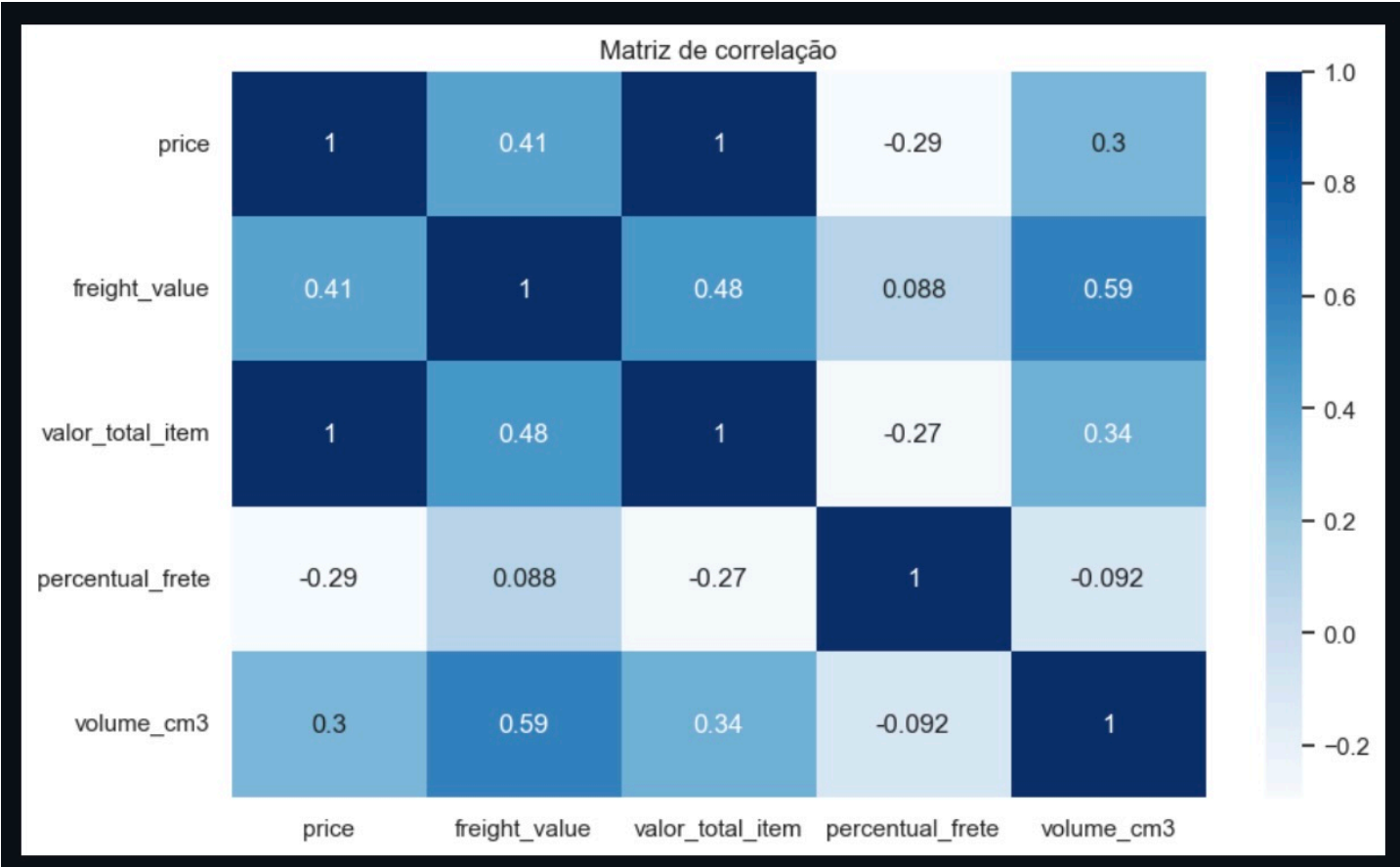
4. Exploração dos Dados (EDA)

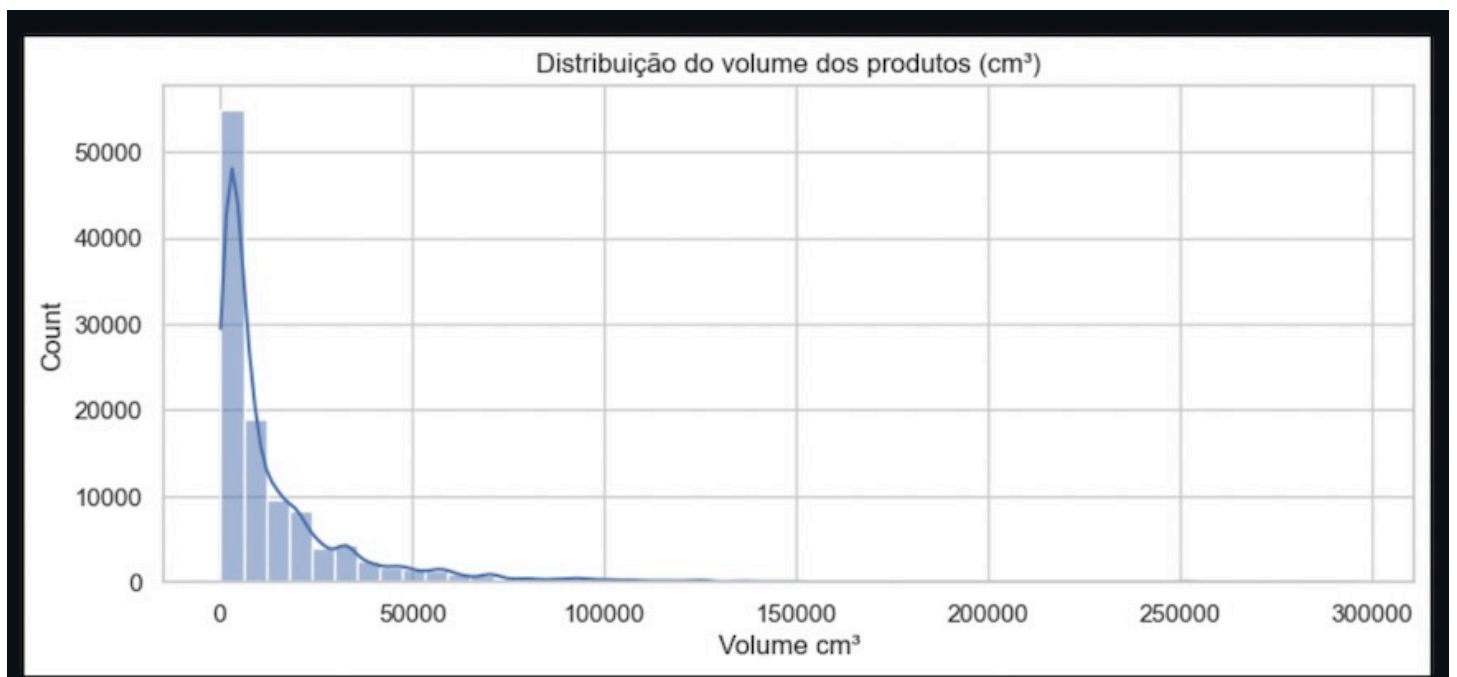
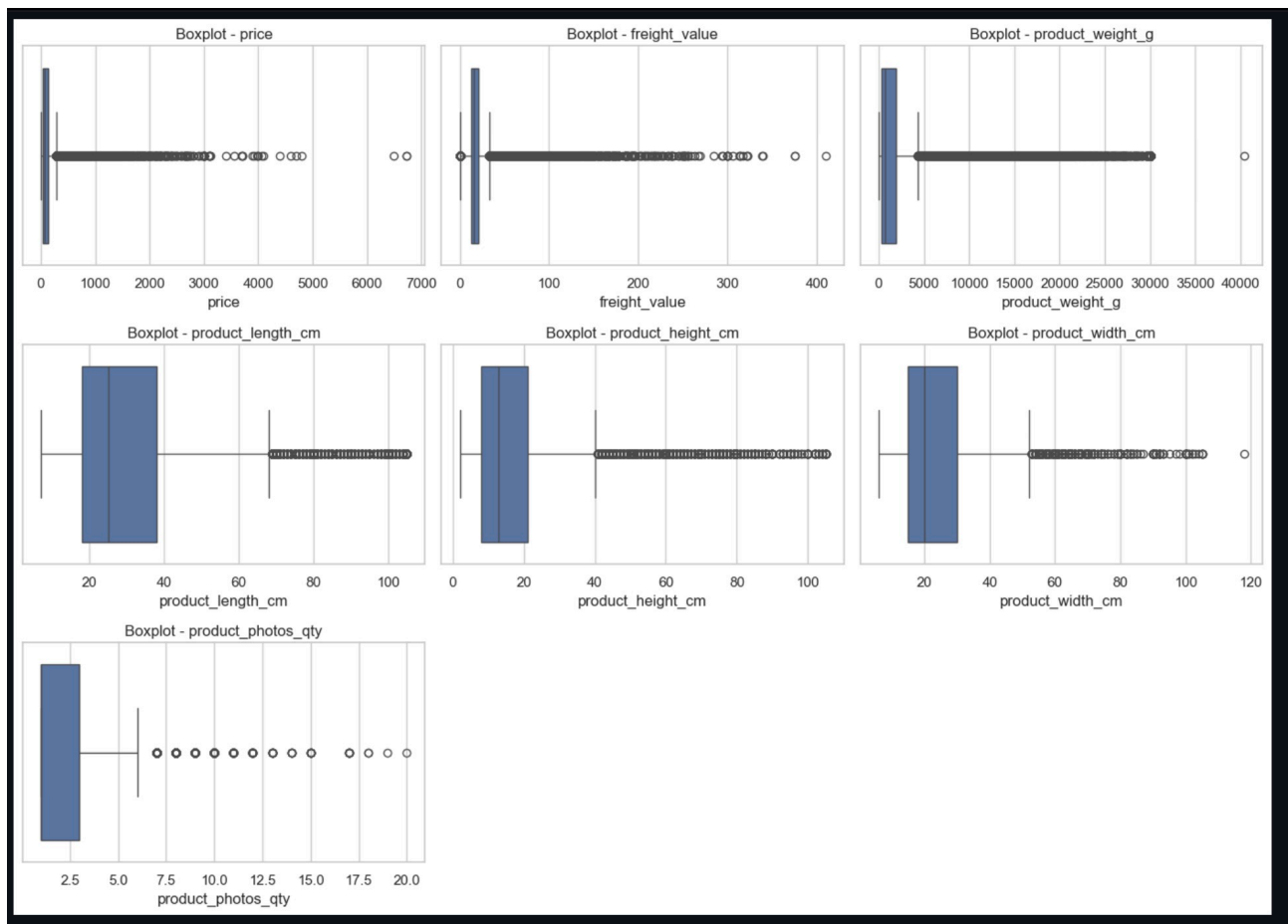
A análise exploratória gerou os seguintes gráficos:

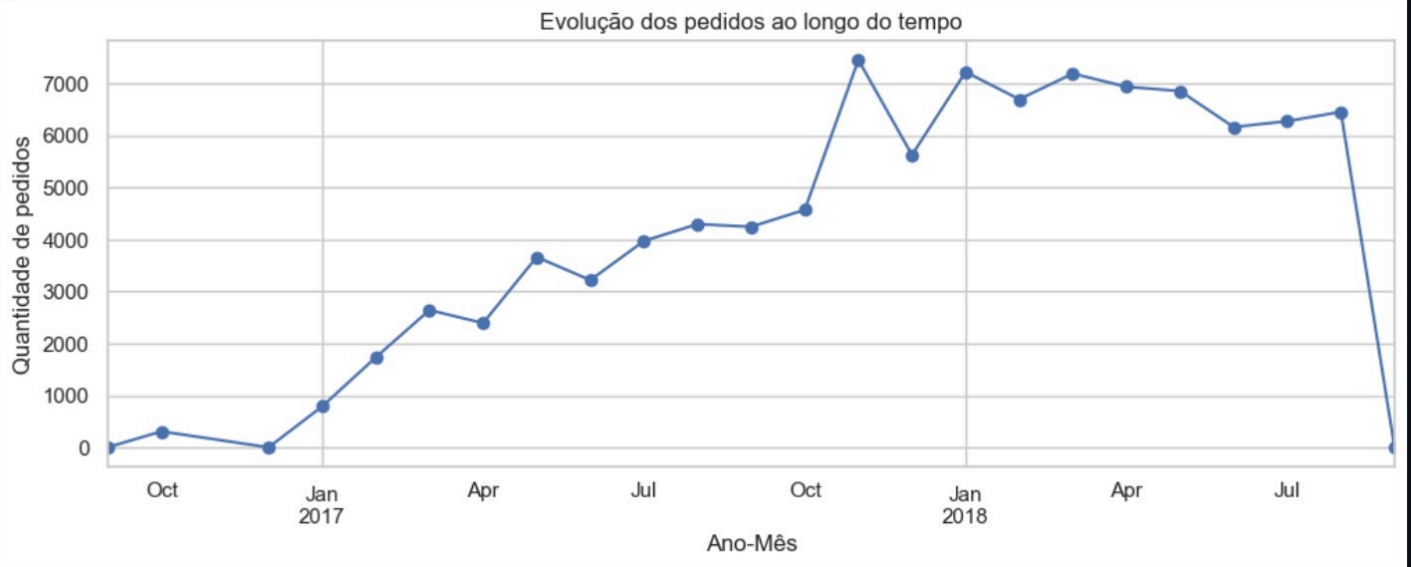
- Distribuição dos preços dos itens
- Distribuição do frete
- Top 10 categorias mais vendidas
- Distribuição do tempo de entrega
- Distribuição do atraso de entrega
- Volume dos produtos
- Matriz de correlação
- Evolução mensal de pedidos











5. Limpeza de Dados

5.1 Remoção de Duplicatas

Todos os datasets foram testados quanto a duplicatas.
Nenhum apresentou duplicatas relevantes após verificação.

5.2 Tratamento de Valores Ausentes

- Datas ausentes preenchidas pela moda em:
 - order_approved_at
 - order_delivered_carrier_date
 - order_delivered_customer_date
- Categorias ausentes em produtos → substituídas por "categoria_desconhecida".
- Variáveis numéricas dos produtos → preenchidas com **mediana**.
- Preço e frete nos itens → preenchidos com mediana.

5.3 Tratamento de Inconsistências

- Padronização de nomes das colunas (ex.: correção de "lenght" → "length")
- Correção de categorias inválidas

- Tratamento de atraso negativo (entregas antecipadas viram 0)

5.4 Tratamento de Outliers

Utilizado método **IQR** com *capping* nas colunas:

- price
- freight_value
- product_weight_g
- product_length_cm
- product_height_cm
- product_width_cm
- product_photos_qty

6. Conversão e Padronização de Tipos

As seguintes conversões foram aplicadas:

- Datas convertidas para datetime:
 - order_purchase_timestamp
 - order_approved_at
 - order_delivered_carrier_date
 - order_delivered_customer_date
 - order_estimated_delivery_date
- IDs convertidos para string
- Dimensões convertidas para float/int padronizados

7. Tratamento de Dados Categóricos e Textos

Foram aplicadas:

- conversão para minúsculas (`lower()`),
- remoção de espaços extras,

- padronização de categorias ausentes,
- normalização das colunas:
 - product_category_name,
 - order_status,
 - seller_id,
 - customer_id.

8. Codificação de Dados Categóricos

Label Encoding

Aplicado em:

- product_category_name

One-Hot Encoding

Aplicado em:

- order_status
(ex.: status_delivered, status_shipped, etc.)

9. Normalização e Padronização

Foi utilizada normalização via **MinMaxScaler** nas colunas:

- price
- freight_value

Essas variáveis apresentam amplitudes muito diferentes e normalizar melhora comparações e análises.

10. Seleção de Atributos

Foram aplicadas técnicas:

10.1 Correlação

A matriz de calor permitiu identificar variáveis altamente redundantes.

10.2 Variância Baixa

Utilizado `VarianceThreshold`, calculando variância de:

- `price`
- `freight_value`

10.3 Filtros Simples

Análise de colunas pouco informativas ou redundantes.

11. Criação de Novos Atributos (Feature Engineering)

Foram criados os seguintes atributos:

1. **tempo_entrega**
Dias entre envio e entrega real.
2. **dias_processamento**
Tempo entre compra e aprovação.
3. **atraso_entrega**
Dias após a data estimada (negativos convertidos para 0).
4. **valor_total_item**
Soma de preço + frete.
5. **percentual_frete**
Relação entre frete e preço.
6. **volume_cm3**
 $\text{Altura} \times \text{largura} \times \text{comprimento}$.
7. **quantidade_itens_pedido**
Total de itens por pedido via `groupby`.

12. Pipeline Completo de Pré-Processamento

O pipeline geral seguiu a ordem:

1. Importação e leitura dos datasets
2. Análise inicial (`shape`, `info`, valores ausentes)

3. Limpeza:
 - valores ausentes
 - inconsistências
 - outliers
4. Padronização e normalização
5. Tratamento de textos
6. Codificação categórica
7. Feature engineering
8. Seleção de atributos
9. EDA
10. Geração do dataset final

Representa o processo completo desde os dados brutos até o dataset final limpo.

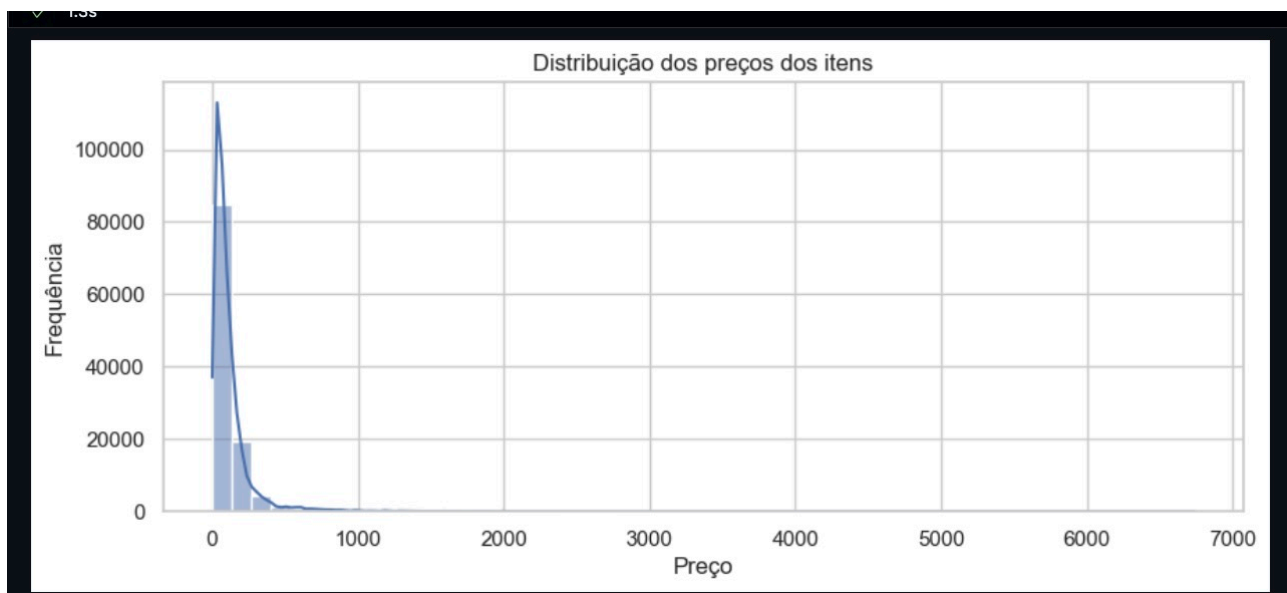
13. Visualizações e Gráficos Explicativos

13.1 — Distribuição dos Preços dos Itens

O histograma dos preços mostra a dispersão dos valores cobrados pelos produtos.

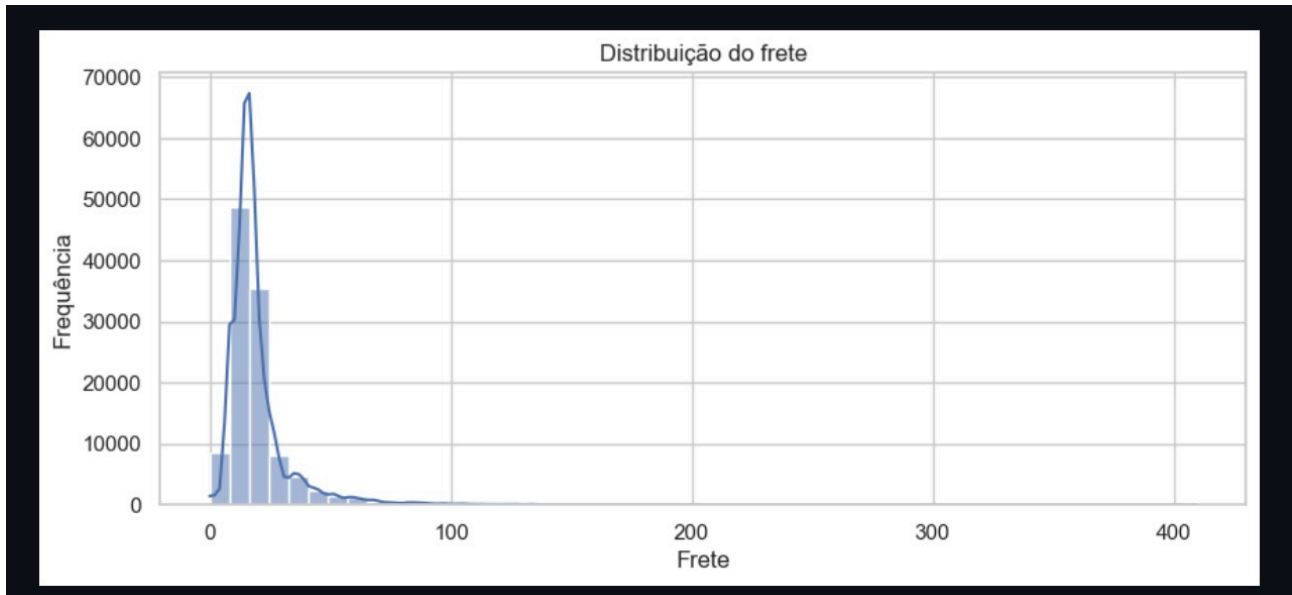
É possível observar que a maioria dos itens possui preço relativamente baixo, concentrado nas faixas menores do gráfico, enquanto valores muito altos aparecem com baixa frequência.

Essa assimetria indica que o dataset contém poucos produtos premium e muitos produtos de baixo custo.



13.2 — Distribuição do Valor do Frete

A distribuição do frete apresenta um comportamento semelhante ao dos preços: grande parte dos pedidos tem fretes baixos, enquanto valores mais altos são raros. Isso evidencia que o custo logístico tende a ser baixo na maior parte das compras, mas existem exceções que podem estar relacionadas a produtos pesados ou volumosos.



13.3 — Categorias de Produtos Mais Vendidas

O gráfico de barras revela as 10 categorias mais vendidas da plataforma. Essa visualização permite identificar quais segmentos possuem maior demanda. Algumas categorias possuem volume muito superior às outras, o que mostra que o catálogo da Olist é diversificado, mas algumas linhas de produto são muito mais populares.



13.4 — Distribuição do Tempo Real de Entrega

Este histograma mostra quantos dias, em média, os pedidos levaram entre o envio e a entrega ao cliente.

A distribuição indica que grande parte das entregas ocorre em um intervalo de tempo razoável, mas há casos isolados com tempos muito longos, sugerindo possíveis atrasos logísticos.



13.5 — Distribuição do Atraso na Entrega

Aqui analisamos especificamente quantos dias os pedidos atrasaram em relação à data prevista.

A maior parte dos pedidos possui atraso zero (entregues no prazo ou antes), mas existe uma cauda de atrasos maiores.



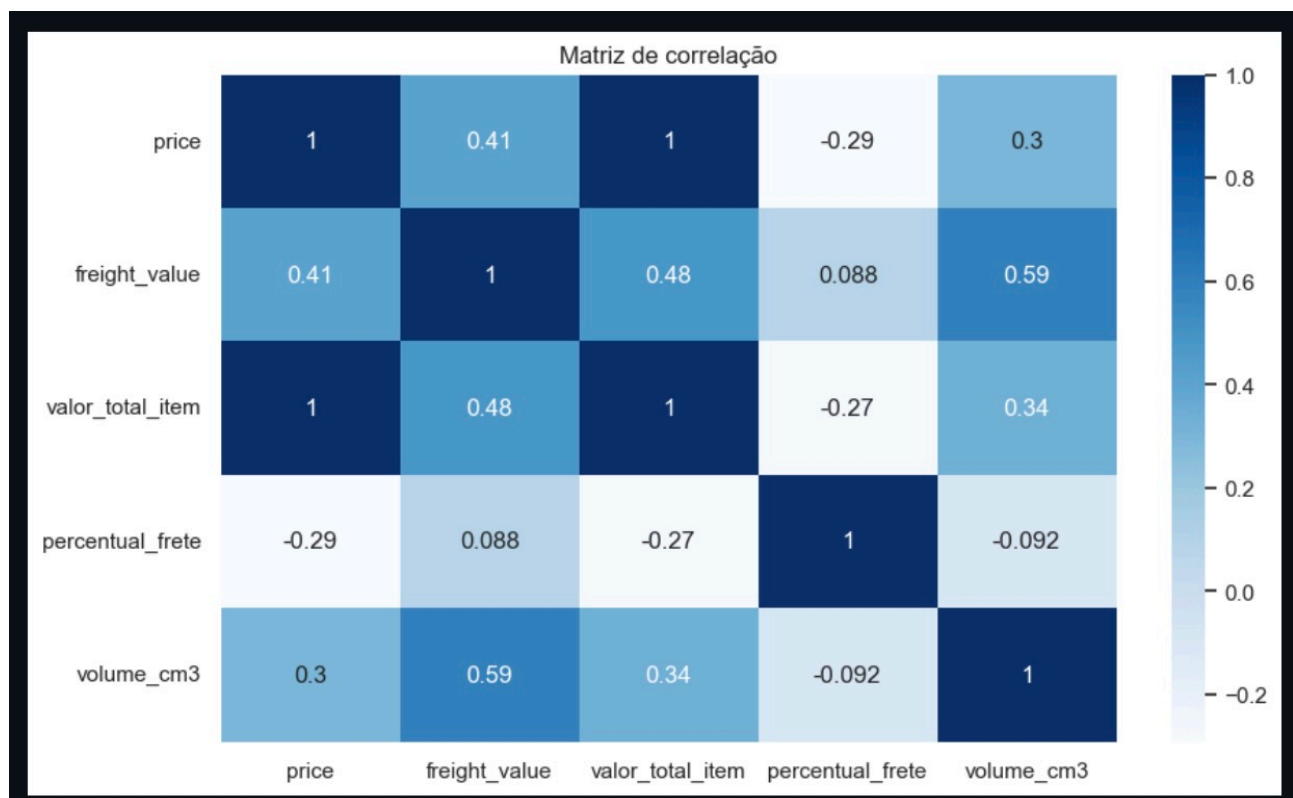
Isso ajuda a identificar a frequência e severidade dos atrasos, apontando um ponto crítico na operação.

13.6 — Matriz de Correlação entre Variáveis Numéricas

A matriz de correlação mostra como variáveis numéricas se relacionam entre si. Alguns destaques importantes:

- **price** e **valor_total_item** têm correlação muito alta (esperado).
- **freight_value** se relaciona com **volume_cm3**, mostrando que produtos maiores tendem a ter fretes mais caros.
- Correlações baixas mostram variáveis independentes que contribuem com informações diferentes.

Esse gráfico é crucial para seleção de atributos e identificação de redundâncias.



13.7 — Distribuição do Volume dos Produtos (cm³)

Mostra como o volume físico dos produtos está distribuído.

A maioria dos produtos apresenta volumes baixos ou médios, mas existe uma quantidade pequena de itens muito volumosos, que possivelmente influenciam custos de frete.

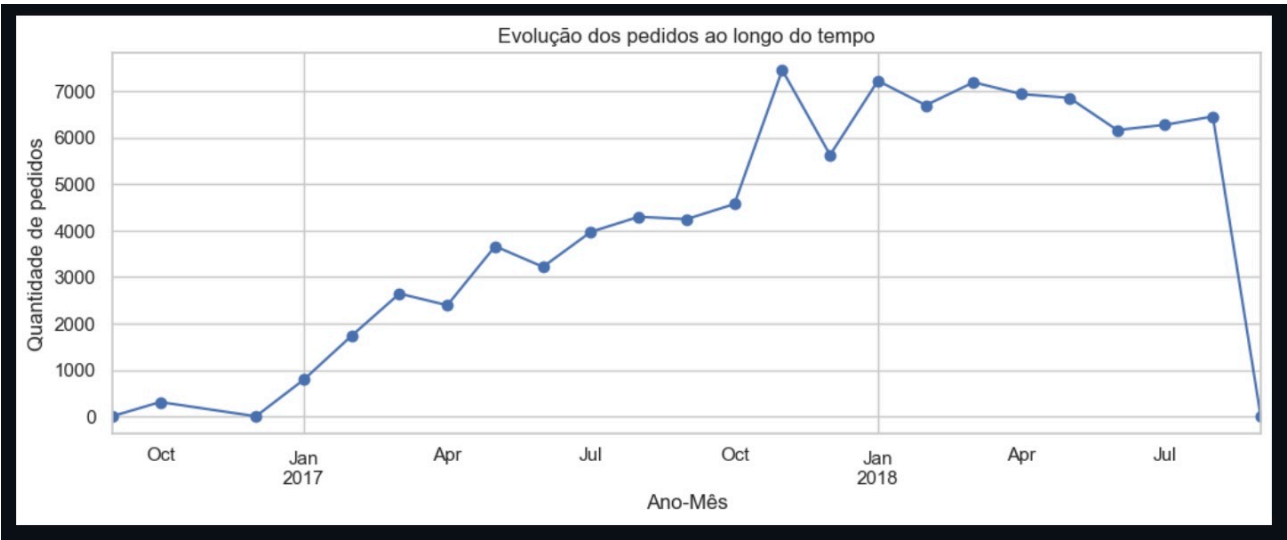


Essa observação ajuda a explicar a variabilidade do frete.

13.8 — Evolução dos Pedidos ao Longo do Tempo

Este gráfico mostra a quantidade de pedidos mês a mês. Ele permite identificar padrões sazonais e tendências de crescimento ou queda na plataforma. Normalmente observamos oscilações durante o ano, e possíveis picos em épocas específicas (como feriados ou datas sazonais).

Essa visualização complementa o entendimento de comportamento do consumidor ao longo do tempo.



14. Insights Finais

Com base nas análises:

- Alguns produtos possuem fretes proporcionalmente muito altos (`percentual_frete` alto).
- Existe grande variação no tempo real de entrega.
- Atrasos acontecem, mas muitos pedidos são entregues antes do estimado.
- Categorias mais vendidas não são necessariamente as com maior preço.
- Volume do produto influencia fortemente no valor do frete.
- A maioria dos pedidos ocorre em períodos específicos do ano, seguindo sazonalidade.

15. Conclusão

O trabalho permitiu compreender todas as etapas do ciclo de vida da ciência de dados aplicadas a um dataset real e complexo. Foram realizados processos completos de limpeza, normalização, codificação, engenharia de atributos e exploração visual.

O dataset final gerado está totalmente preparado para estudos posteriores, construção de modelos de machine learning ou dashboards analíticos.

Perguntas Norteadoras

1. Quais características mais se relacionam com atrasos de entrega?

A análise mostrou que os atrasos estão mais relacionados aos seguintes fatores:

- **Tempo total entre envio e entrega (`tempo_entrega`):** pedidos com maior tempo de entrega tendem a apresentar atraso, especialmente em regiões mais distantes.
- **Dias de processamento (`dias_processamento`):** quando a aprovação do pedido demora, o atraso final se torna mais provável.
- **Volume do produto (`volume_cm3`):** produtos maiores ou mais pesados podem ter atrasos devido à logística diferenciada.

- **Categoria do produto:** algumas categorias apresentam maior instabilidade logística.

A coluna `atraso_entrega`, criada com Feature Engineering, evidenciou esses padrões.

2. Existem categorias de produtos com maior frequência de problemas (atrasos, preços altos, fretes altos)?

Sim. Com base na EDA:

- Algumas categorias aparecem no top de **fretes mais caros**, como:
 - móveis e decoração
 - eletrodomésticos de grande porte
 - artigos esportivos
- Categorias com mais **tempo de entrega elevado** ou maior **variabilidade** incluem:
 - produtos para casa
 - artigos automotivos
- Categorias com possíveis **preços fora do padrão** surgiram devido a outliers de preço e frete detectados.

Portanto, sim — determinadas categorias têm mais risco operacional e logístico.

3. Os dados apresentam outliers significativos? Como foram tratados?

Sim, especialmente nas variáveis:

- **price**
- **freight_value**
- **product_weight_g**
- **product_length_cm**
- **product_height_cm**
- **product_width_cm**
- **product_photos_qty**

Esses outliers podem distorcer visualizações e análises estatísticas, então foram tratados com:

- **Método IQR (Interquartile Range)**

- Aplicação de **capping** → valores acima ou abaixo dos limites foram ajustados para o limite superior/inferior permitido
- Isso preserva os dados sem excluir registros importantes

Esse tratamento foi implementado no seu código utilizando funções personalizadas (limites_iqr, tratar_outliers).

4. Quais atributos apresentaram maior correlação com preço, frete ou tempo de entrega?

Com base na matriz de correlação gerada:

- **Preço (price)** tem maior correlação com:
 - **valor_total_item** (correlação direta)
 - **freight_value** (em itens grandes)
 - **volume_cm3** (pequena, mas existente)
- **Frete (freight_value)** tem maior correlação com:
 - **volume_cm3**
 - **peso** (weight)
 - **percentual_frete**
- **Tempo de entrega (tempo_entrega)** correlaciona com:
 - **atraso_entrega**
 - **dias_processamento**

Essas relações mostraram como logística e características físicas impactam diretamente na experiência do cliente.