

**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.

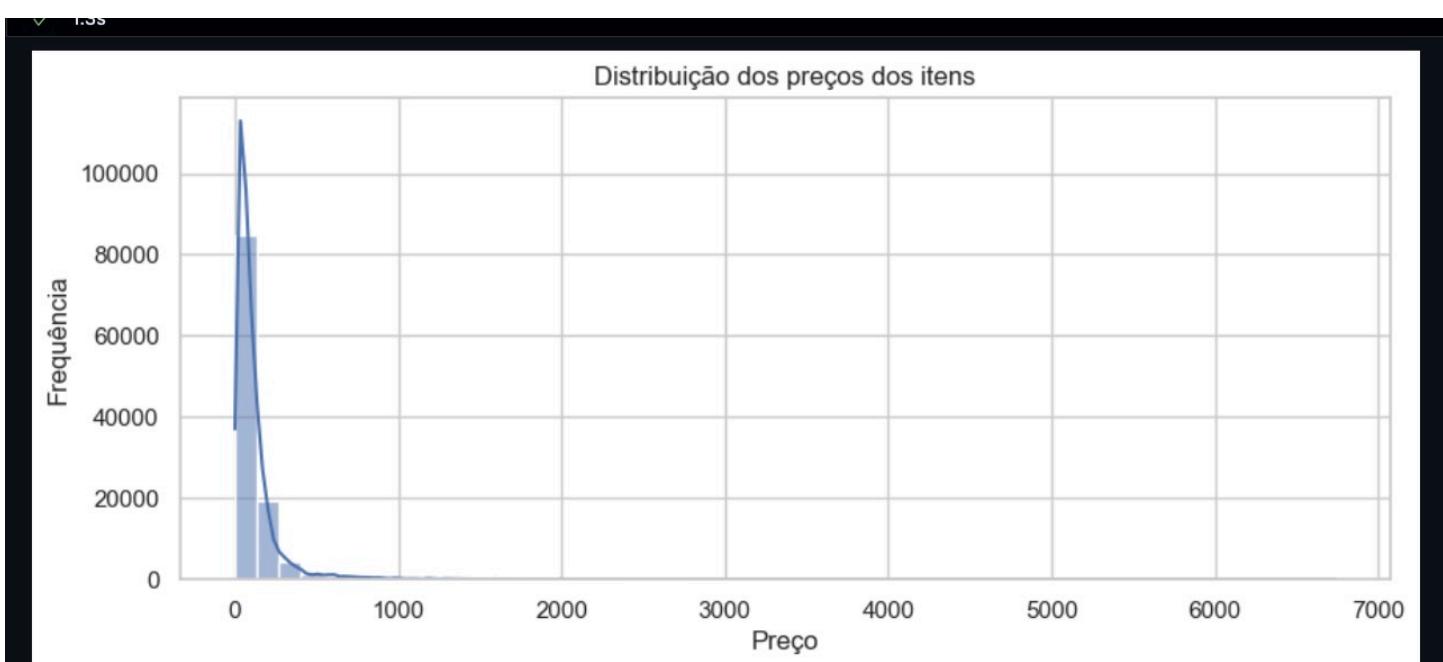
### 3.6 Comunicação dos Resultados

Geração de insights, gráficos e conclusões.

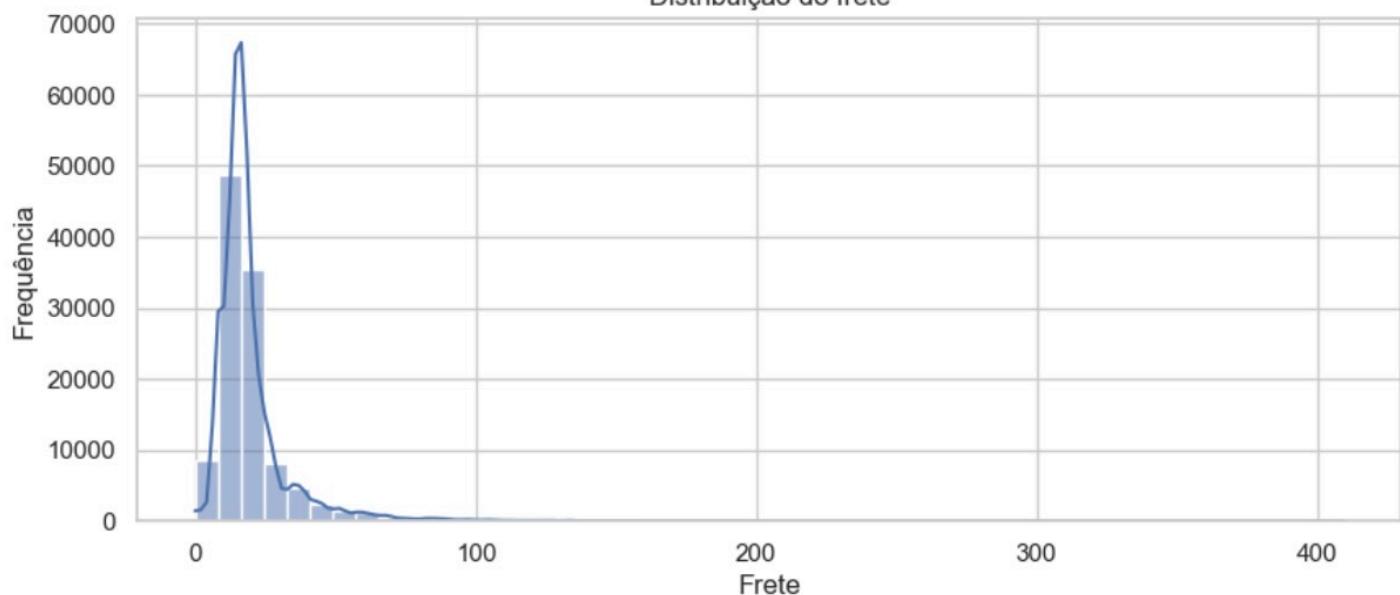
## 4. Exploração dos Dados (EDA)

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

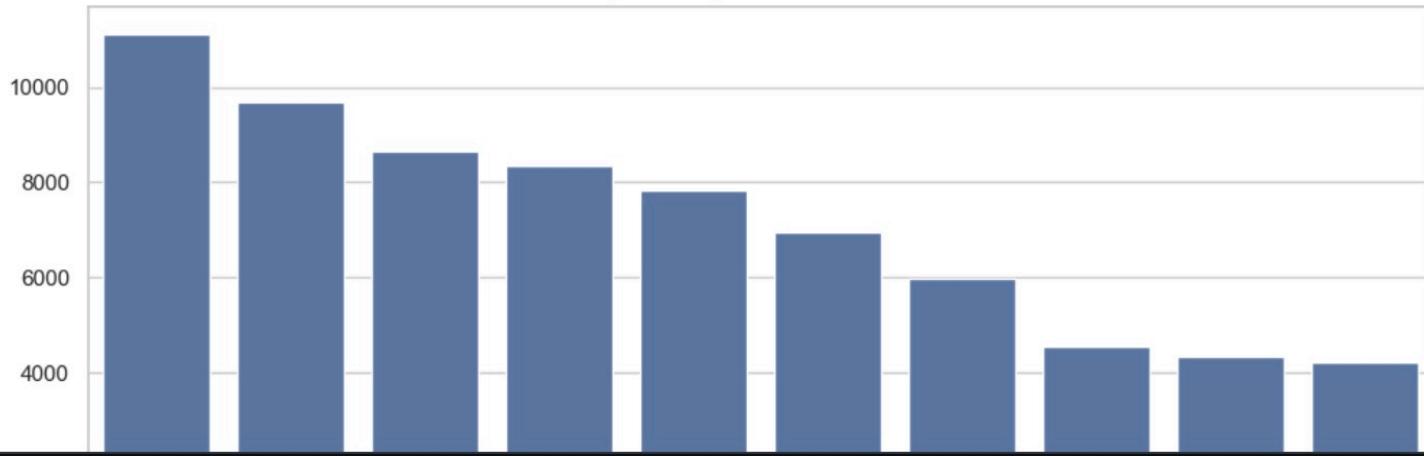


Distribuição do frete



Top 10 categorias mais vendidas

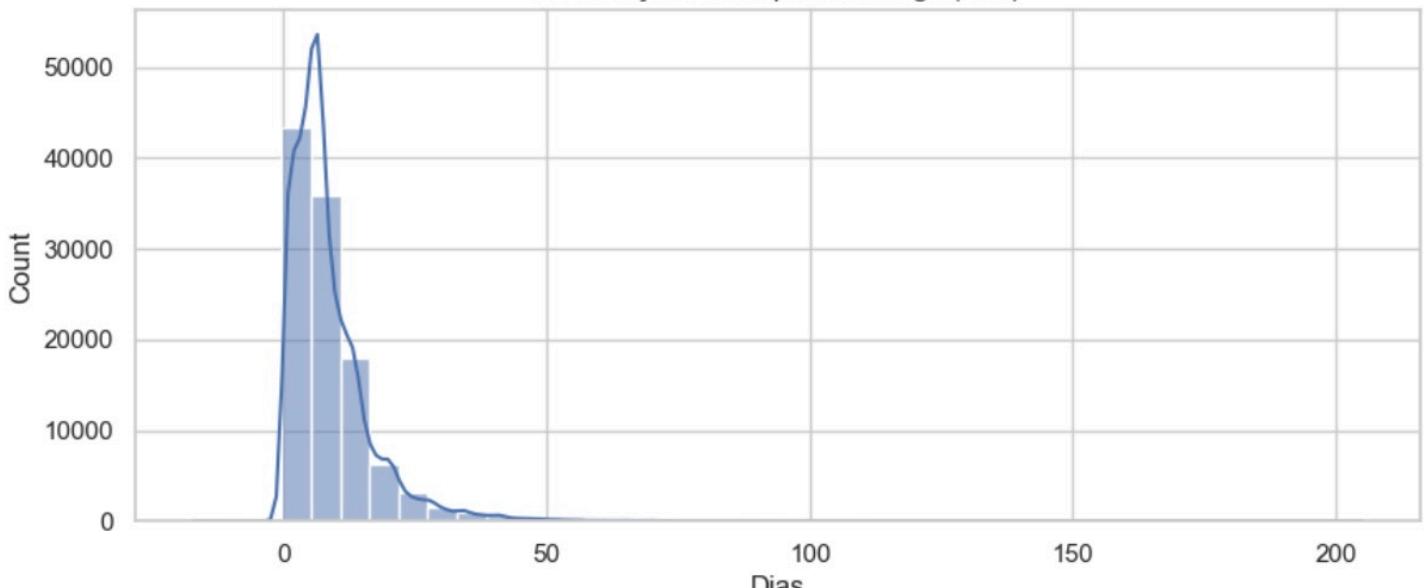
Quantidade vendida



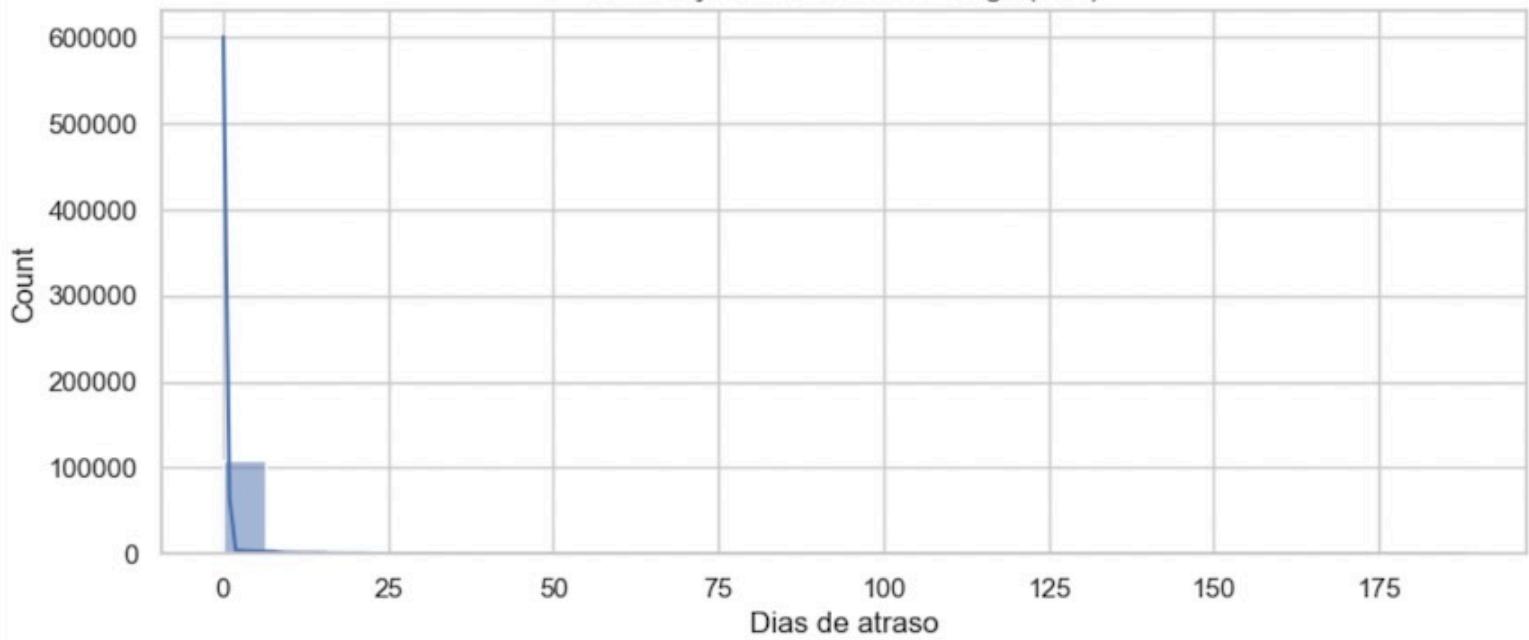
Distribuição do tempo de entrega (dias)

Count

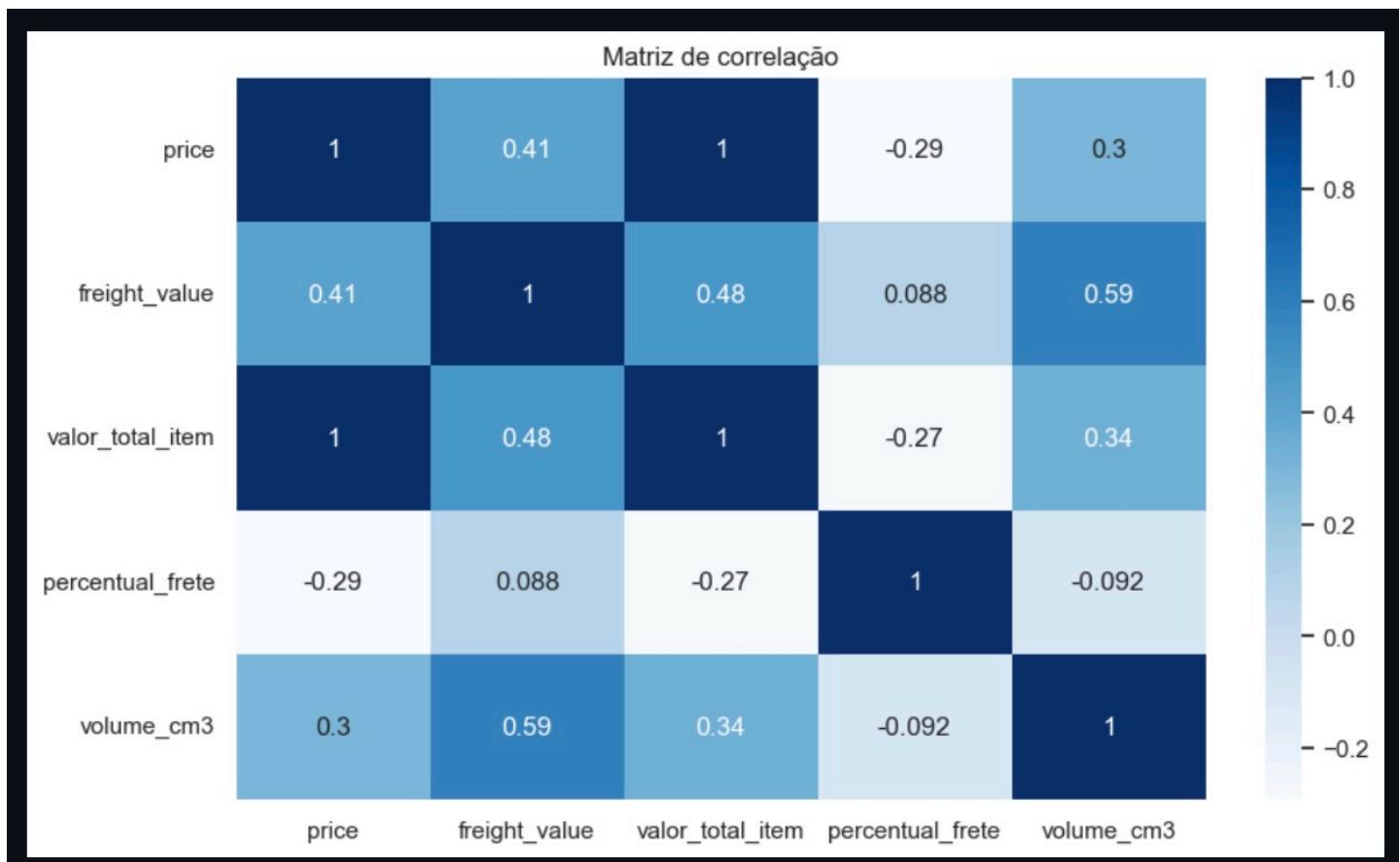
Dias

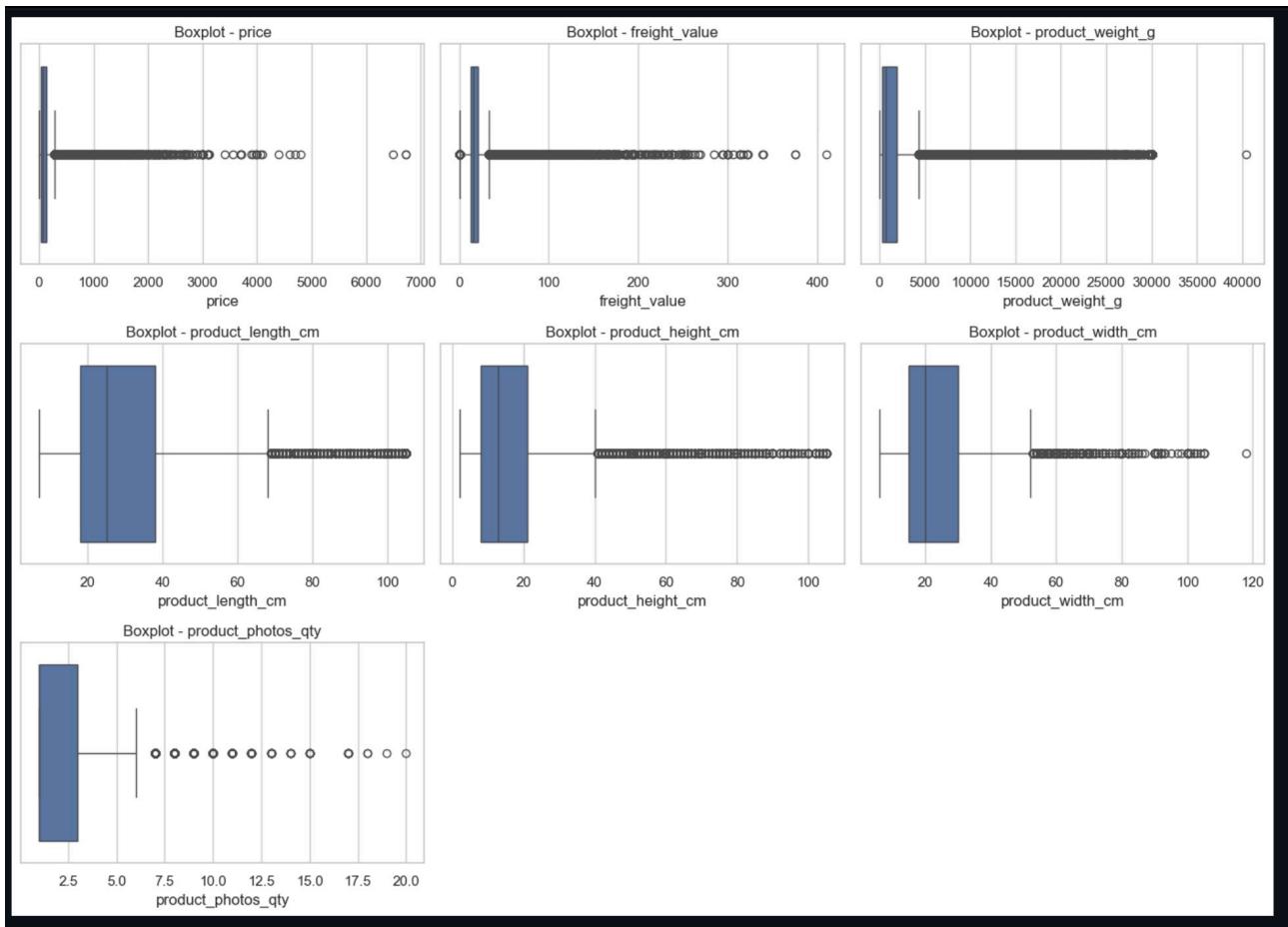


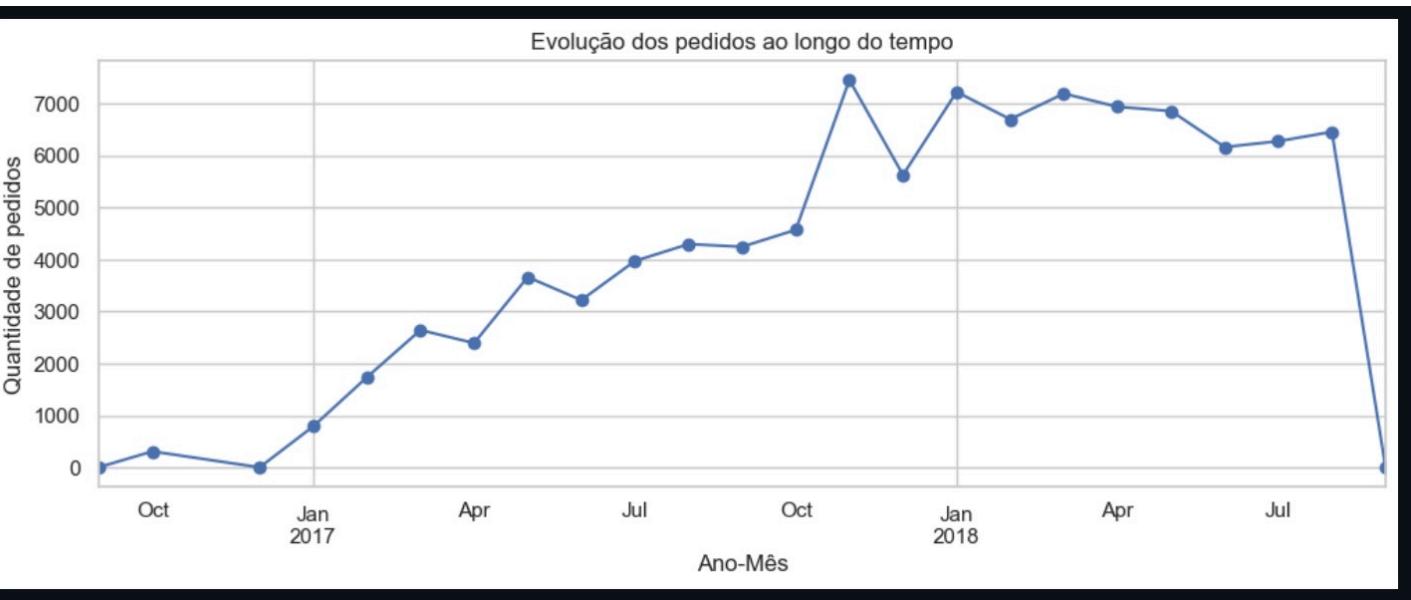
Distribuição do atraso de entrega (dias)



Matriz de correlação







## 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**  
Altura × largura × 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

Aqui devem ser colocados:

- histogramas de preço, frete, volume
- barras das categorias mais vendidas
- evolução mensal de pedidos
- gráficos de atraso e tempo de entrega
- matriz de correlação

Cole as imagens geradas no notebook.

## 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.