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

Cidade – Teresina
Ano (2025)

1. Contextualização do Problema

O projeto analisa dados reais do e-commerce brasileiro (Olist) para entender fatores que influenciam:

- atrasos de entrega
- custos (preço e frete)
- qualidade dos produtos
- volume de vendas
- impacto das características dos produtos no desempenho logístico

O objetivo é aplicar o ciclo completo de Data Science, explorando, limpando, transformando e preparando os dados para análises e modelos futuros.

2. Bases de Dados Utilizadas

Fonte: *Brazilian E-Commerce Public Dataset by Olist*
Arquivos usados:

- **olist_orders_dataset.csv**
- **olist_order_items_dataset.csv**
- **olist_products_dataset.csv**

Cada dataset representa uma etapa do processo de compra:

- **Orders:** informações gerais do pedido e datas.
- **Order Items:** preço, frete e quantidade por item.
- **Products:** características físicas dos produtos.

3. Ciclo de Vida da Ciência de Dados Aplicado

Etapas executadas:

1. **Coleta dos dados** (arquivos CSV da Olist)
2. **Entendimento dos dados** (tipos, formatos, tamanhos)
3. **Limpeza e tratamento** (NAs, duplicatas, outliers)
4. **Transformações** (tipos, escalonamento, codificação)
5. **Feature Engineering**

6. **EDA — Análises e gráficos**
7. **Geração do dataset final pré-processado**

4. Exploração Inicial (EDA)

Principais verificações:

- Distribuição de **preços**, **frete**, **volume**
- Frequência de categorias
- Evolução temporal de pedidos
- Correlação entre variáveis numéricas

As visualizações ajudaram a identificar a necessidade de padronização, remoção de valores extremos e criação de novos atributos.

5. Limpeza dos Dados

5.1 Duplicatas

- Verificação realizada nos três datasets.
- Não foram encontradas duplicatas relevantes.

5.2 Inconsistências

Principais problemas detectados:

- Colunas com nomes errados (product_description_lenght).
- Produtos com medidas igual a **0** (altura, peso etc.).
- Datas ausentes ou incoerentes.

Correções foram aplicadas conforme o dataset.

5.3 Valores Ausentes

Orders

- Datas faltantes preenchidas com **moda** (valor mais recorrente).

Products

- Categorias nulas → "categoria_desconhecida"
- Variáveis numéricas → preenchidas com **mediana**

Items

- price e freight_value → preenchidos com mediana.

5.4 Outliers

Método aplicado:

- Cálculo de limites via **IQR**
- Substituição por limites inferiores/superiores

Colunas tratadas:

- Preço
- Frete
- Medidas físicas
- Volume

6. Conversão e Padronização de Tipos

- Colunas de data convertidas para datetime.
- IDs tratados como strings.
- Variáveis categóricas normalizadas para caixa baixa.
- Medidas convertidas para tipo numérico.

7. Tratamento de Dados Categóricos e Textos

Procedimentos:

- Normalização de nomes de categorias
- Preenchimento de ausentes
- Padronização de strings
- Transformação de categorias raras em "outras" (quando aplicável)

8. Codificação de Variáveis Categóricas

Codificações utilizadas:

- **One-Hot Encoding:** categorias de produto

- **Label Encoding:** colunas com alta cardinalidade (se necessário)

9. Normalização e Padronização

Técnicas aplicadas:

- **MinMaxScaler:** para modelos sensíveis a distância
- **StandardScaler (Z-score):** para variáveis com distribuição normalizada

Variáveis transformadas:

- Preço
- Frete
- Volume
- Peso

10. Seleção de Atributos

10.1 Correlação

Atributos analisados:

- price
- freight_value
- valor_total_item
- percentual_frete
- volume_cm3

10.2 Baixa Variância

Variáveis quase constantes foram removidas ou ignoradas.

10.3 Filtros Simples

Atributos irrelevantes excluídos:

- IDs
- timestamps não utilizados em modelos

11. Feature Engineering (4+ Criadas)

Atributos criados:

1. **valor_total_item** (price × quantidade)
2. **percentual_frete** (frete / preço)
3. **volume_cm3** (altura × largura × comprimento)
4. **tempo_entrega** (entrega → pedido)
5. **atraso_entrega** (dias além da data estimada)
6. **ano_mes** (para análise temporal)

12. Pipeline Completo de Pré-Processamento

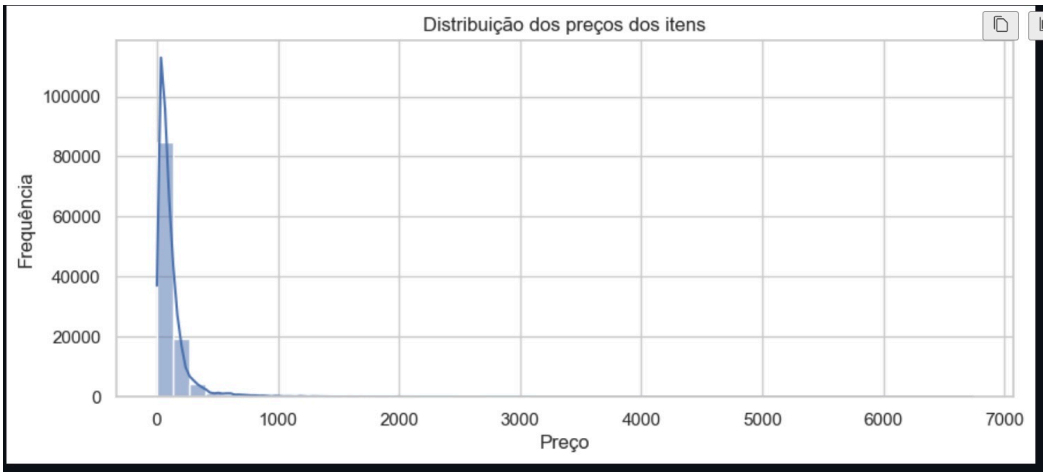
Pipeline aplicado:

1. Conversão de tipos
2. Limpeza de valores ausentes
3. Padronização de categorias
4. Correção de inconsistências
5. Tratamento de outliers
6. Feature engineering
7. Codificação
8. Normalização / Padronização
9. Geração do dataset final

13. Visualizações e Gráficos

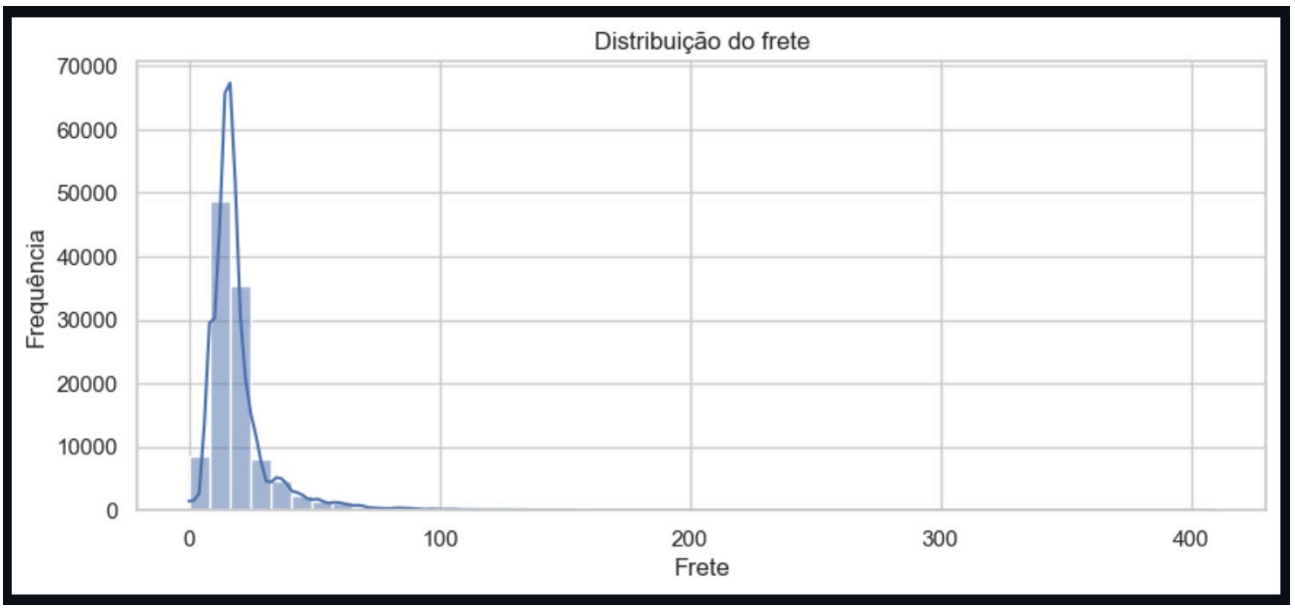
Gráficos presentes:

- Distribuição de preços

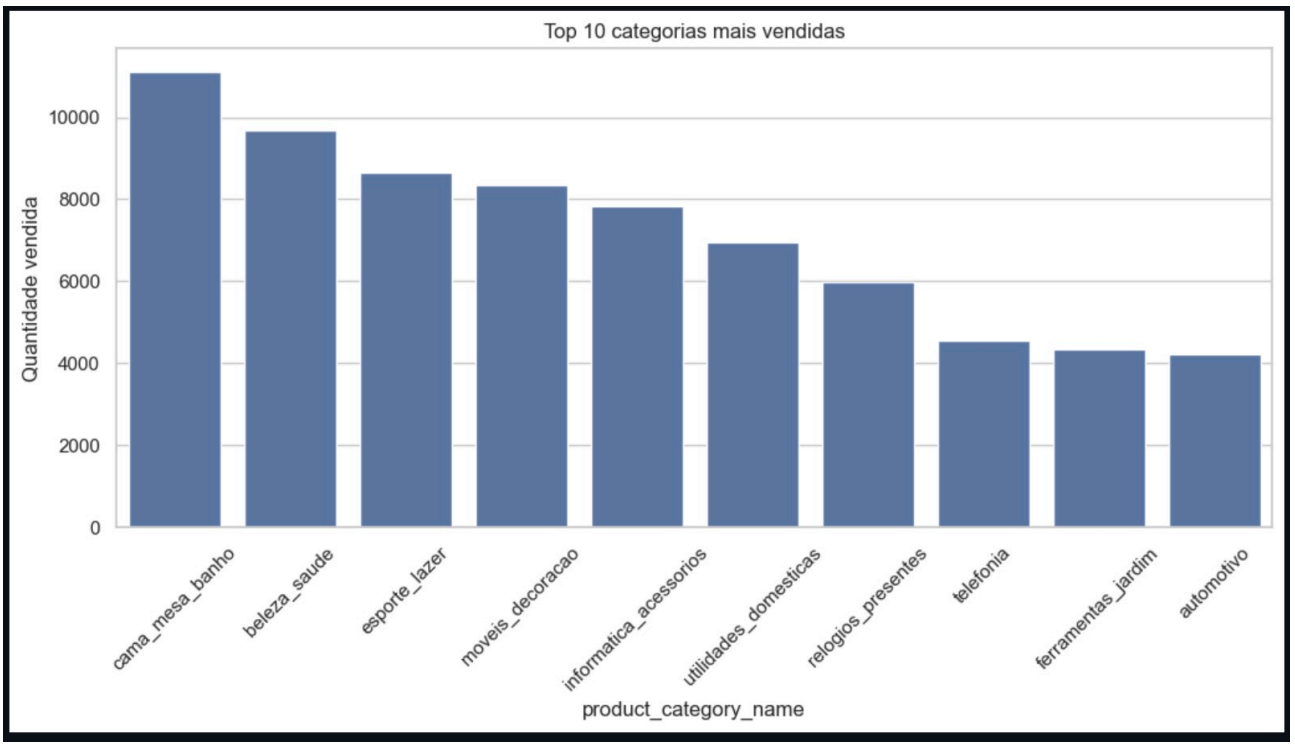


-

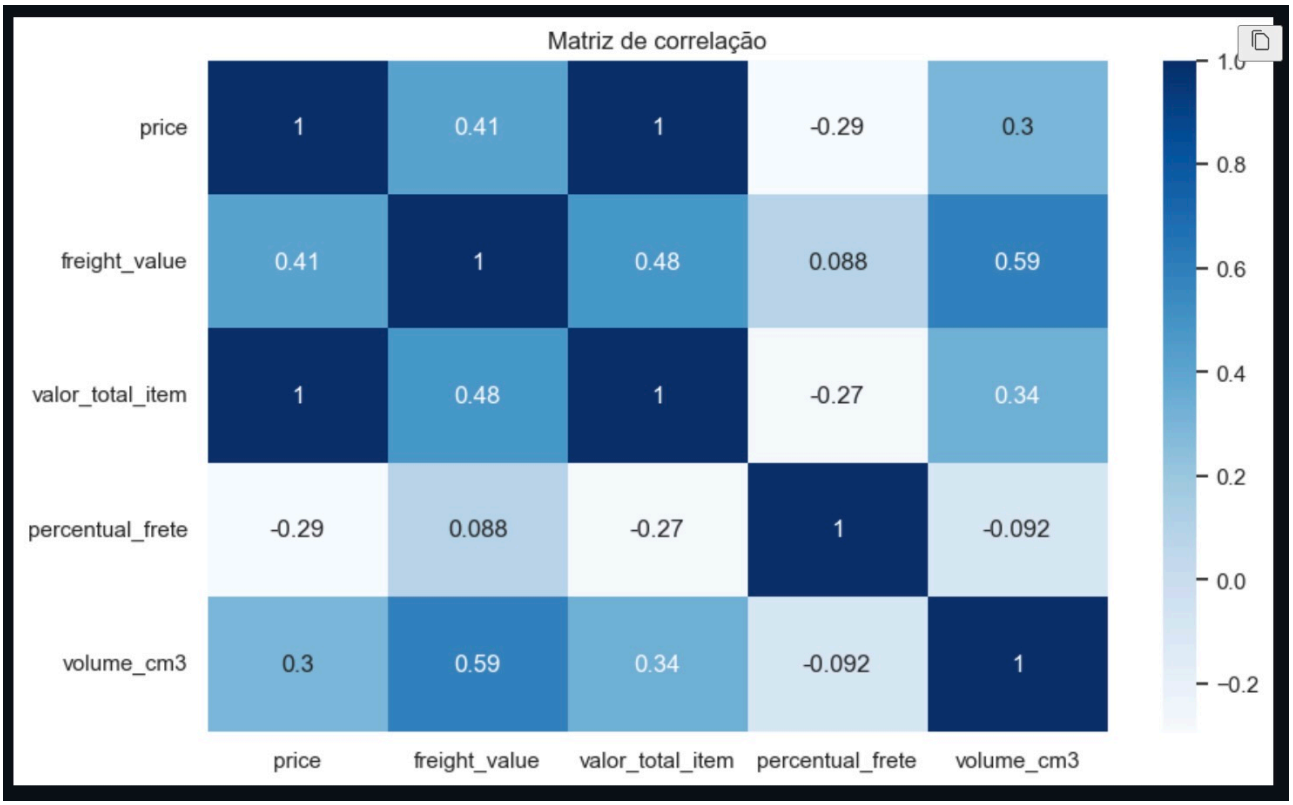
- Distribuição de frete



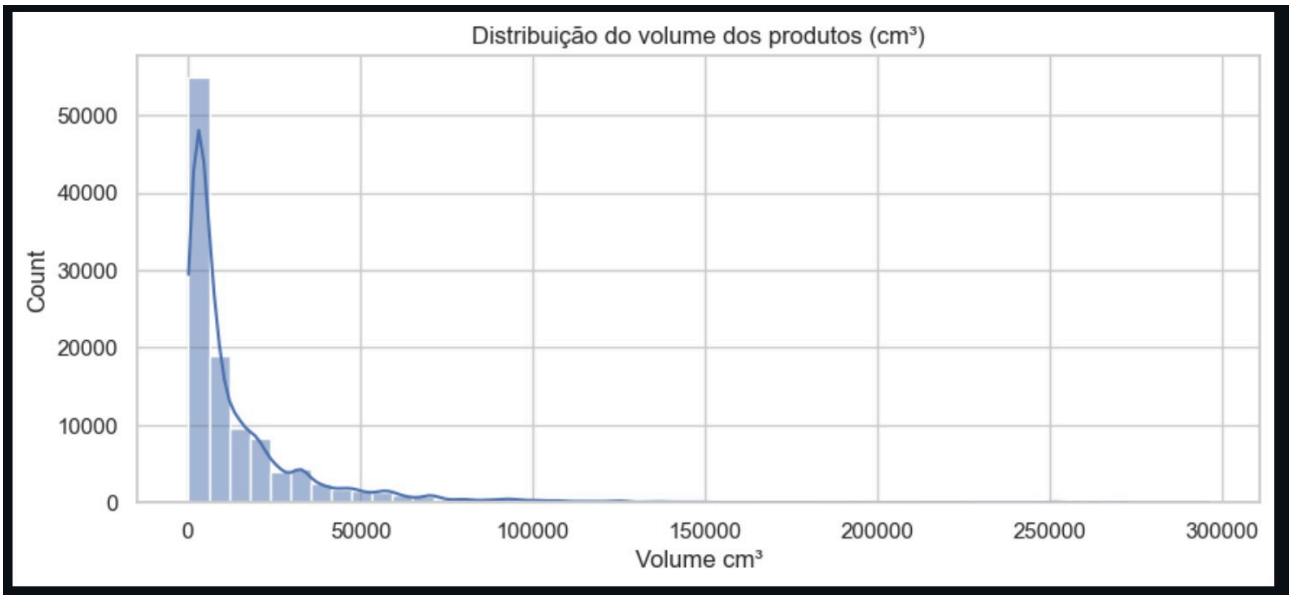
- Top 10 categorias



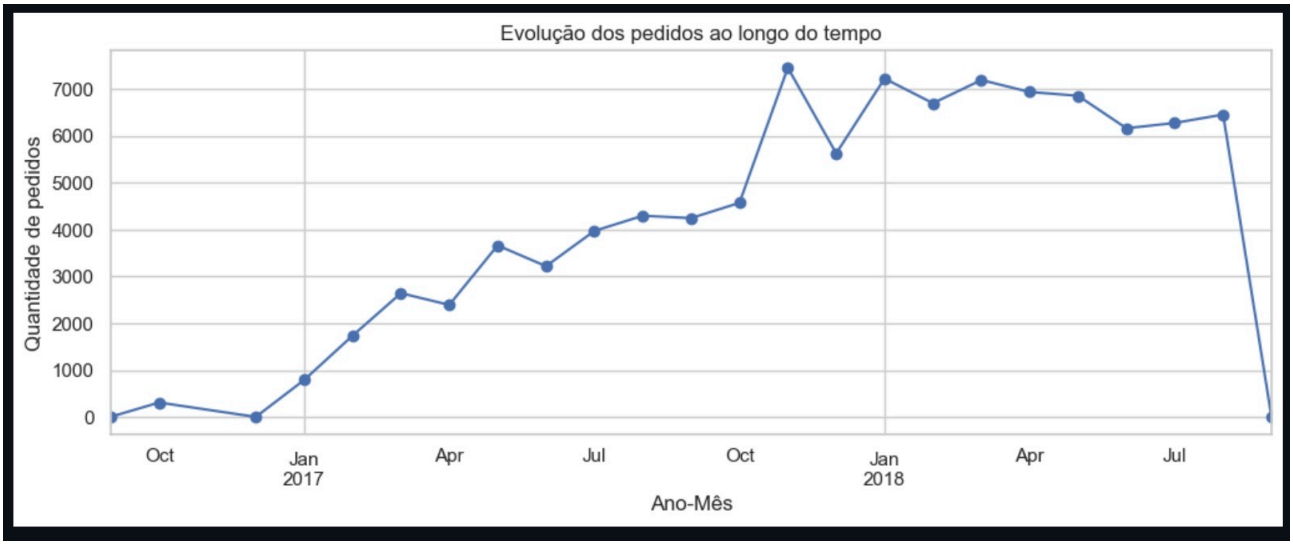
- Correlação



- Volume

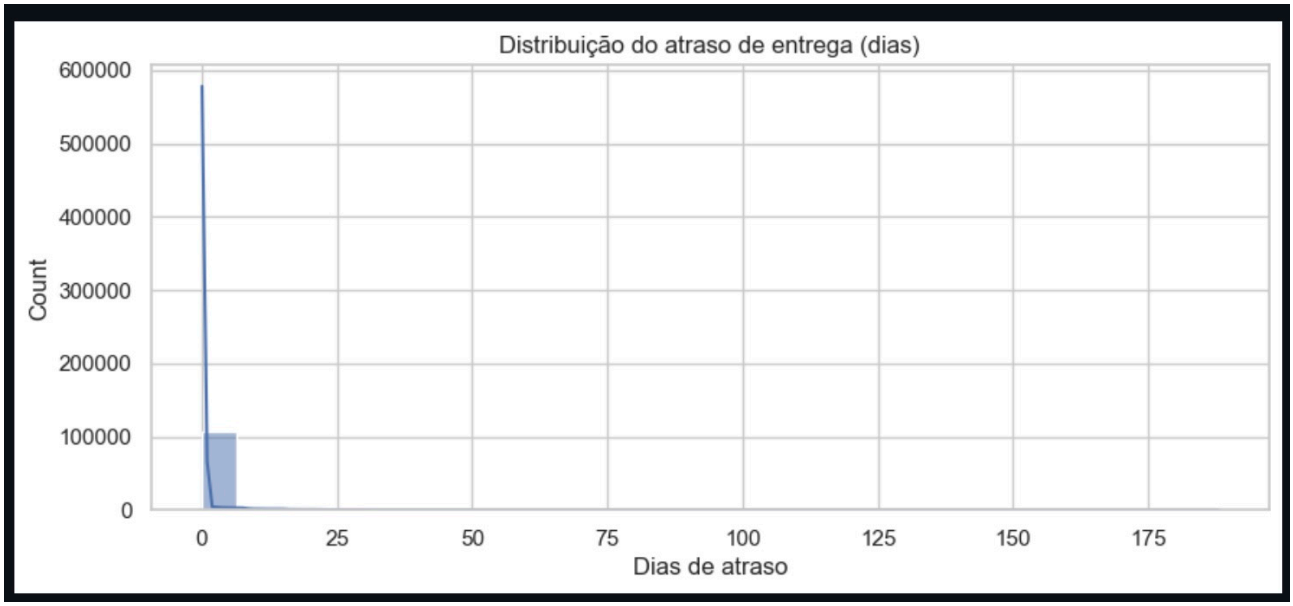


- Evolução de pedidos por mês



Distribuição do tempo de entrega (dias)

Distribuição do atraso de entrega (dias)



14. Insights Finais

Principais descobertas:

- Produtos maiores e mais pesados tendem a **gerar fretes mais altos**.
- Categorias específicas concentraram mais atrasos (ex.: móveis).
- O volume e peso do produto explicam boa parte do custo logístico.
- Atrasos estão ligados ao tempo entre aprovação e envio.
- O frete pode representar até **40% do preço** em alguns produtos.

15. Conclusão

O pré-processamento permitiu limpar, padronizar e transformar os dados, revelando relações importantes entre características dos produtos, logística e atrasos.

Perguntas Norteadoras Respondidas

1) Quais características mais se relacionam com atraso de entrega?

- Maior tempo entre *compra → aprovação*.
- Maior volume e peso do produto.
- Categorias com logística mais lenta (móveis, eletrônicos grandes).

2) Existem categorias com maior frequência de problemas?

Sim — categorias volumosas, móveis e itens grandes apresentam mais frete alto e atrasos.

3) Existem outliers significativos? Como foram tratados?

Sim — preços, fretes e medidas físicas tinham valores extremos.
Foram tratados com limites via **IQR**.

4) Quais atributos mais correlacionam com preço e frete?

- **Preço ↔ valor_total_item**
- **Frete ↔ volume_cm3**
- **Frete ↔ peso**
- **Frete ↔ percentual_frete**