CENTRO UNIVERSITÁRIO SANTO AGOSTINHO - UNIFSA
CURSO DE ENGENHARIA DE SOFTWARE – 4º PERÍODO
DISCIPLINA: CIÊNCIA DE DADOS
PROFESSOR: HELOISA GUIMARÃES COELHO

TRABALHO FINAL – CIÊNCIA DE DADOS

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

ALUNOS:
ANTONIO NEVES AGUIAR NETO
WICKHAM CARNEIRO PEREIRA

Teresina - PI
24.11.2025

CONTEXTUALIZAÇÃO DO PROBLEMA REAL

Impulsionado pela facilidade de compra e diversidade de produtos, o comércio eletrônico brasileiro está bastante presente no dia a dia da população, em contrapartida, apresenta vários desafios logísticos significativos, como atrasos na entrega, inconsistências nos cadastros e grande variação de características de produtos. Essas imperfeições são típicas de um ambiente real de comercio online. Diante desse cenário, surge o problema real tratado nesse trabalho: O que afeta a experiência e satisfação do cliente no e-commerce brasileiro?

APRESENTAÇÃO DOS DATASETS ADOTADOS

O dataset principal utilizado neste trabalho faz parte de um conjunto de 9 datasets disponibilizados pela Olist, um ecossistema de soluções voltado à automatização da gestão de empresas em todo o Brasil. Esse conjunto reúne informações de aproximadamente 100 mil pedidos realizados entre 2016 e 2018, provenientes de diferentes marketplaces. Esses dados permitem analisar aspectos como status do pedido, preço, frete, logística, características dos produtos, entre outros. No entanto, para esse trabalho vamos utilizar apenas os 3 solicitados:

Olist_order_items_dataset – Itens vendidos em cada pedido

Esse dataset relaciona cada pedido aos produtos comprados, representa apenas compras reais, por isso foi usada como tabela principal no processo de merge, para evitar pedidos incompletos e cancelados.

Tem atributos como: order_id, Product_id, seller_id, order_item_id, price, freight_value, shipping_limit_date

Olist_orders_dataset – Informações dos pedidos

Esse dataset tem os dados de cada pedido realizado na plataforma, tem atributos como: order_id, custumer_id, order_status e datas do processo.

Olist_products_dataset – Informações dos produtos

Esse dataset tem atributos de cada produto vendido, tem atributos como: Product_id, Product_category_name, Product_name_lenght, Product_description_lenght, Product_photos_qty. Esse dataset em específico possui vários valores ausentes, o que veremos depois.

APLICAÇÃO DO CICLO DE VIDA DA CIÊNCIA DE DADOS

Aplicamos o ciclo de vida durante o desenvolvimento de todo o trabalho

Business Understanding (Compreensão do Negócio)

Começamos entendendo o problema central e pesquisamos os principais motivos que afetam a experiencia e satisfação do cliente no e-commerce, encontramos relação com atrasos na entrega e diferença de peso/frete, além de variações no tempo de processamento e envio.

Data Understanding (Compreensão dos Dados)

Examinamos os três datasets principais e obrigatórios do trabalho (orders, order_items e products), analisando estrutura, colunas, porcentagem de valores ausentes, outliers e relações entre as tabelas.

Utilizamos comandos como print de colunas, df.head(), df.info(), df.describe(), e um heatmap para ver as correlações.

Data Preparation (Preparação dos Dados)

Incluiu limpeza de duplicatas, padronização textual, conversão de tipos, imputação de valores ausentes (mediana e categoria “Desconhecido”), tratamento de outliers via IQR, integração das tabelas por chaves e criação de novos atributos logísticos, como volume, densidade, atraso e tempo de aprovação.

Modeling (Modelagem)

Começamos entendendo o problema central e pesquisamos os principais motivos que afetam a experiencia e satisfação do cliente no e-commerce, encontramos relação com atrasos na entrega e diferença de peso/frete, além de variações no tempo de processamento e envio.

Evaluation (Avaliação)

Após aplicar o pré-processamento completo, realizamos uma avaliação dos resultados obtidos para verificar a qualidade do tratamento dos dados. Foram avaliados pontos como: Redução e correção de inconsistências, comparação das distribuições antes e depois da limpeza, impacto na imputação de valores ausentes, validação da coerência dos novos atributos criados, verificação visual por meio de gráficos.

Deployment (Deploy)

Realizamos a entrega dos resultados e do notebook contendo todas as etapas do ciclo de vida da Ciência de Dados e pipeline de pré-processamento. Organização do notebook (colunas, seções e descrições), disponibilização do dataset final tratado com documentação detalhada de cada célula e suas funções e enviamos o relatório final contendo toda a metodologia e insights extraídos.

EXPLORAÇÃO DOS DADOS (EDA)

A primeira coisa que realizamos antes de começar os processos de alteração nos datasets foi a exploração inicial dos dados, realizamos procedimentos como:

- Mostrar todas as colunas de cada uma das tabelas;
- Mostrar as 5 primeiras linhas do dataset já unificado;
- Mostrar as informações gerais sobre o dataframe;
- Mostrar as informações descritivas sobre o dataframe;
- Mostrar um Heatmap para ver as correlações de algumas colunas.

Todos esses procedimentos foram realizados para entender melhor os dados que iriamos trabalhar, verificar se o merge havia funcionado da maneira ideal e observar as correlações das colunas, durante esse procedimento começamos a ter os primeiros insights.

Percebemos que algumas colunas deveriam estar em outro formato, como por exemplo: order_purchase_timestamp,order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_estimated_delivery_date, shipping_limit_date. Essas colunas deveriam estar como datetime.
Temos outras colunas que deveriam estar em formatos diferentes, vamos concertar tudo isso na seção de conversão e padronização de tipos

LIMPEZA DE DADOS

NULOS

Primeiro resolvemos tratar os nulos, decidimos que os nulos nas colunas "product_name_lenght", "product_photos_qty" e "Product_description_lenght" será colocado com valor 0, futuramente no trabalho converteremos valor para INT, porque quando se tem valores NaN o Dtype sempre volta como object, então para fazer a conversão depois, imputamos NAN como 0;

Na coluna product_category_name, valores nulos será colocado "Desconhecido" para ter uma maneira de nomear esses valores nulos.

Depois, vamos pegar as colunas "product_weight_g","product_length_cm", "product_height_cm", "product_width_cm", onde for nulo vamos fazer a mediana da categoria para substituir, pois mediana é menos afetada por outliers.

Verificamos vários valores nulos nas colunas relacionadas a data também, mas não iremos imputar valores ainda, pois o objetivo é que esses valores ausentes virem NaT após a conversão para datetime.

DUPLICATAS

Em relação a duplicatas, percebemos que elas podem existir na maioria das colunas sem nenhum problema, com exceção dessas duas colunas juntas: “order_id” e “order_item_id”

Fizemos um print para ver se essas colunas tinham duplicatas, mas não encontramos nenhuma.

INCONSISTÊNCIAS

Olhando o df.describe, percebemos que algumas coisas possuem peso 0, o que é uma inconsistência lógica, então buscamos além dessa categoria, se existem valores 0 ou negativos em: “Price”, “Freight_value” e “Product_Weight_g”.

Porque em preço, nada deve ter valor negativo ou zerado, em valor do frete não deve haver fretes negativos, mas podemos ter fretes zerados e em peso do produto não devemos ter peso negativo nem zerados.

Acabamos encontrando 8 valores inconsistentes no peso do produto, decidimos tratar eles usando mediana para substituir valores inconsistentes, pois mediana não é distorcida por outliers.

Depois colocamos um código para tratar possíveis inconsistências por mapeamento no nome da categoria, com o objetivo de evitar categorias duplicadas acidentalmente, seja por espaços diferentes ou letras maiúsculas\minúsculas.

OUTLIERS

Depois de observar os valores dos outliers das colunas preço, valor do frete e peso do produto encontramos algumas porcentagens interessantes, 7.48% da coluna price, 10.77% da coluna freight_value e 14.03% da coluna Product_weight_g são outliers. Mas decidimos não mexer nesses outliers pois durante o projeto trabalhamos com valores reais e optamos apenas por tirar medidas irreais.

Modificamos também as colunas de tamanho, largura e comprimento, que possuem valor 0 pois eles destoam do valor. (Esses valores zerados não são nulos, eles já estavam com zero antes, por isso não tratamos ele na aba de nulos) Então substituímos valores menores ou igual a 0 pela mediana da categoria.

CONVERSÃO E PADRONIZAÇÃO DE TIPOS

Identificamos que as seguintes colunas precisavam ser convertidas: Todas as que possuem data e as Product_description_lenght, Product_photos_qty e Product_name_lenght.

As que possuem data convertemos para datetime e as outras citadas nos parágrafos anteriores foram convertidas para int64, temos que realizar essa conversão porque caso as que são data continuarem como string não poderemos calcular diferença entre datas nem realizar outras operações, já as product_description_lenght, product_photos_qty, product_name_lenght foram convertidos para int64 pois representam valores numéricos, se continuassem como float ou object, não conseguiríamos fazer análises estatísticas ou construir gráficos corretamente.

TRATAMENTO DE DADOS CATEGÓRICO E TEXTOS

Como dito na sessão Limpeza de Dados Inconsistências, para garantir consistência nas análises envolvendo variáveis categóricas, realizamos uma padronização das colunas textuais. Primeiramente, aplicamos str.lower() e str.strip() na coluna product_category_name para eliminar diferenças de caixa (maiúsculas/minúsculas) e remover espaços desnecessários, evitando a criação de categorias duplicadas por pequenos erros de escrita.

CODIFICAÇÃO DE DADOS CATEGÓRICOS

A coluna order_status foi transformada através de codificação categórica, utilizando o método One-Hot Encoding. Esse método é usado para converter cada categoria textual (*“shipped”*, *“delivered”*, *“canceled”*) em colunas binárias contendo valores 0 ou 1. É necessário porque o computador não interpreta texto, apenas valores numéricos. Também evita que o algoritmo intérprete categorias como se houvesse uma relação de ordem entre elas, o que aconteceria com codificação numérica simples.

Escolhemos a coluna order_status por apresentar poucas categorias e possuir relevância direta nas análises logísticas. A aplicação dessa técnica garante que o dataset esteja adequado para as próximas etapas.

FEATURE ENGINEERING - CRIAÇÃO DE NOVOS ATRIBUTOS

Criamos quatro features essenciais relacionadas à logística, atraso e custos, que julgamos estarem mais relacionados aos padrões que tema do projeto pede:

Feature 1 - Delivery_delay_days

Essa feature calcula quantos dias um pedido foi entregue antes, no prazo ou depois da data prevista. Criamos essa feature pois atrasos impactam consideravelmente na satisfação do cliente. Essa métrica revela falhas logísticas e gargalos.

Feature 2 – is_late_delivery

Essa feature converte a coluna para numérica, buscando melhorar gráficos e matrizes de correlação, podendo saber facilmente se a encomenda está atrasada ou não. Criamos pela facilidade ao identificar de forma direta se atrasou ou não e porque permite comparar percentuais de atraso.

Feature 3 – processing_time_days

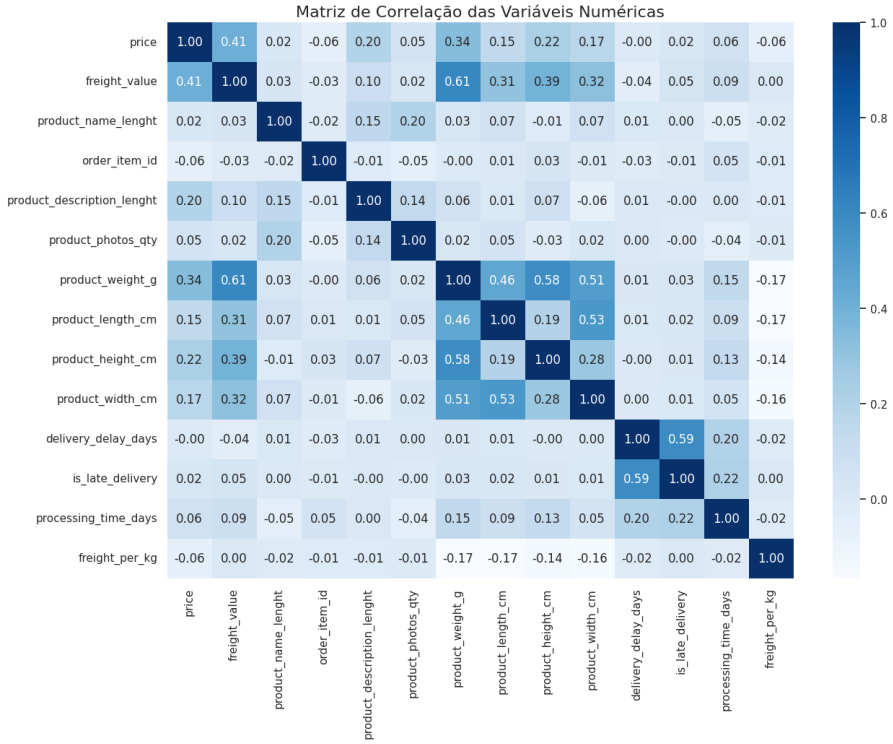
Essa feature serve para ver o tempo em dias entre compra e envio à transportadora, é o tempo que o vendedor leva até entregar o pedido a transportadora. Criamos essa feature pois o desempenho logístico não é responsabilidade apenas da transportadora, muitos atrasos podem ocorrer pois o vendedor demora para despachar o pedido. Assim podemos separar atraso do vendedor e atraso do transporte e ver quais vendedores demoram mais para processar os pedidos e o impacto disso.

Feature 4 – freight_per_kg

Feature criada para ter uma métrica de custo do frete proporcional ao peso, podemos identificar produtos com peso semelhante pode ter frete muito diferente, permitir a identificação de produtos com frete caro demais para o peso real e ver anomalias, como produtos leves com frete muito alto ou produto pesado com frete muito baixo. Também podemos analisar quais produtos tem pior custo-benefício logístico etc.

SELEÇÃO DE ATRIBUTOS

Seleção por correlação

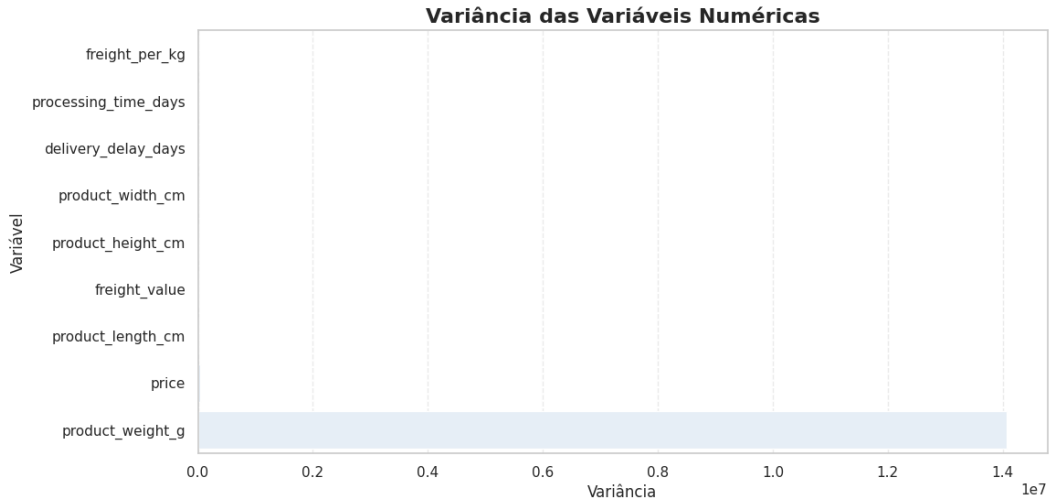


A seleção de atributos por correlação serve para descobrir quais colunas numéricas tem relação, para gerar insights melhores. Primeiro usamos um heatmap para identificar e observamos que:

1. Produtos mais pesados possuem frete maior
2. Produtos maiores causam mais atraso
3. O tempo de processamento afeta o atraso final
4. O frete por quilo está coerente com o peso e o preço

Seleção por baixa variância

A seleção de atributos por baixa variância serve para descobrir variáveis que nunca mudam e, portanto, não servem para gerar insights. Utilizamos um Barplot para melhor visualização da variância:



Com isso observamos que:

1. Coluna "order_item_id" não possui variância e, portanto, não tem poder explicativo;
2. Coluna "is_late_delivery" Não possui variância, mas significa que a maioria dos pedidos não atrasaram;
3. Coluna "freight_per_kg" não possui variância, mas indica que não há inconsistências e nem problemas logísticos;
4. Grande variedade de peso de produtos.

Seleção por Relevância Lógica

Excluímos da normalização e padronização a coluna "order_items_id" e colunas object, pois, ou não fazem sentido em análise, pois não tem variância e correlação válida ou não são usados por conta do seu tipo.

NORMALIZAÇÃO E PADRONIZAÇÃO (MINMAX, Z-SCORE)

Quando normalizamos tudo para mesma escala, conseguimos interpretar padrões logísticos com mais clareza, isso permite identificar por exemplo, produtos com medidas irregulares, fretes desproporcionais ao seu peso, tempos de processamento muito altos ou atrasos incomuns.

Temos a comparação das três versões, na ordem a Original, Minmax e Z-Score. MinMax deixa tudo entre 0 e 1, Z-Score centraliza os valores e deixa os outliers mais visíveis. Colocamos as imagens dos histogramas na parte “visualizações e gráficos explicativos”. Os histogramas mostram que mesmo com as transformações a distribuição se mantém, portanto julgamos que apenas deixa melhor de se comparar atributos e não muda o resultado.

PIPELINE COMPLETO DE PRÉ PROCESSAMENTO

O pipeline de pré-processamento corresponde a toda a sequência de etapas aplicadas aos dados antes das análises finais. Ele abrange o processo completo de preparação, desde a inspeção inicial até a normalização dos atributos numéricos. Realizamos nesse trabalho a seguinte pipeline de pré-processamento:

1. **Carregamento e exploração inicial**

   Examinamos estrutura, tipos e estatísticas com df.head(), df.info() e df.describe().

2. **Identificação dos valores ausentes**

   Usamos df.isnull().sum() para contabilizar nulos e compreender quais colunas exigiam tratamento.

3. **Tratamento dos valores ausentes**

   Aplicamos imputação colocando a mediana por categoria em variáveis numéricas;

   Colunas categóricas com valores ausentes receberam a categoria “desconhecido”.

4. **Correção de inconsistências**

Ajustamos pesos iguais a zero ou negativos usando a mediana por categoria;

Padronizamos textos com str.lower() e str.strip() para evitar duplicação de categorias.

5. **Codificação das variáveis categóricas**

   Utilizamos One-Hot Encoding (pd.get_dummies) na coluna order_status.

6. **Normalização/Padronização das variáveis numéricas**

   Aplicamos StandardScaler para padronizar atributos numéricos e evitar distorções na análise.

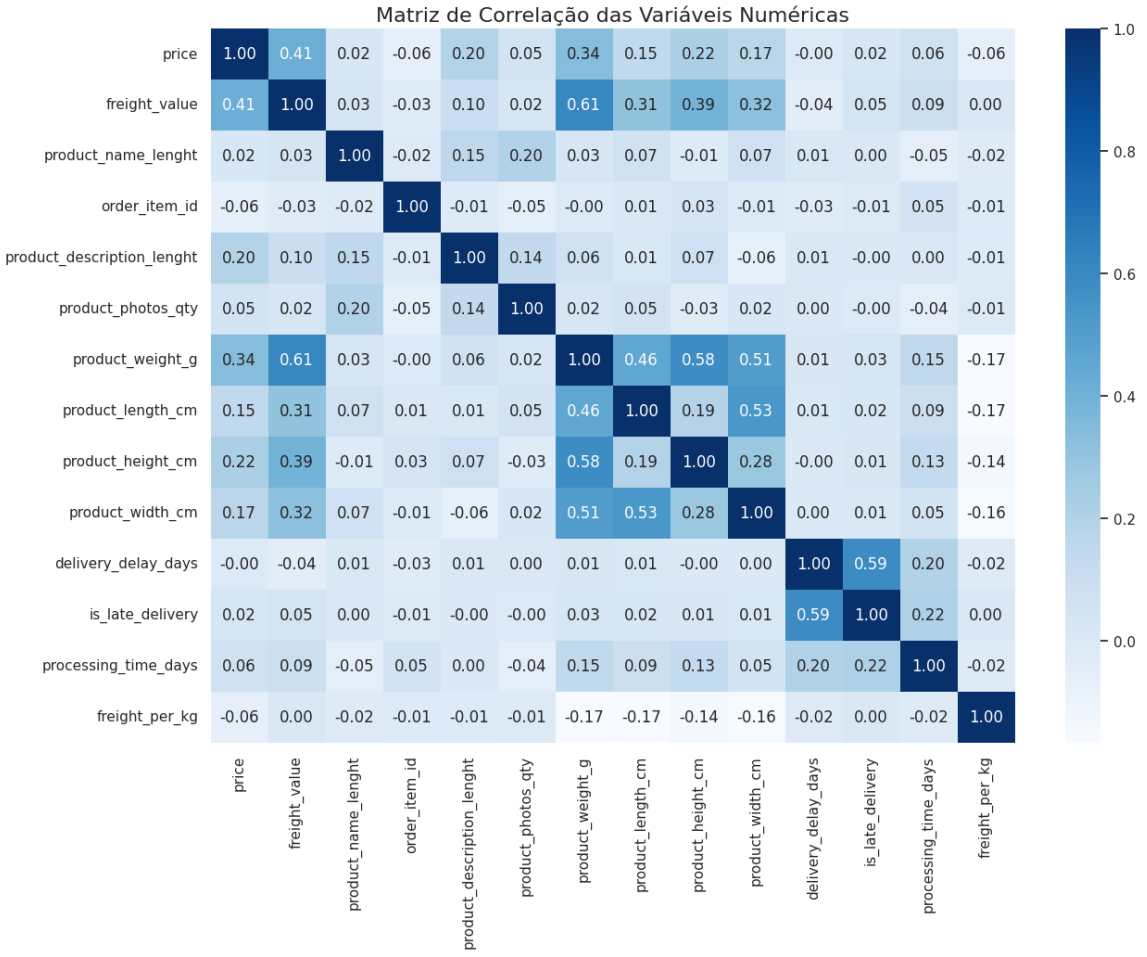
7. **Criação de novas features (Feature Engineering)**

   Criamos atributos como tempo de entrega, atraso, densidade, volume e custo logístico por peso.

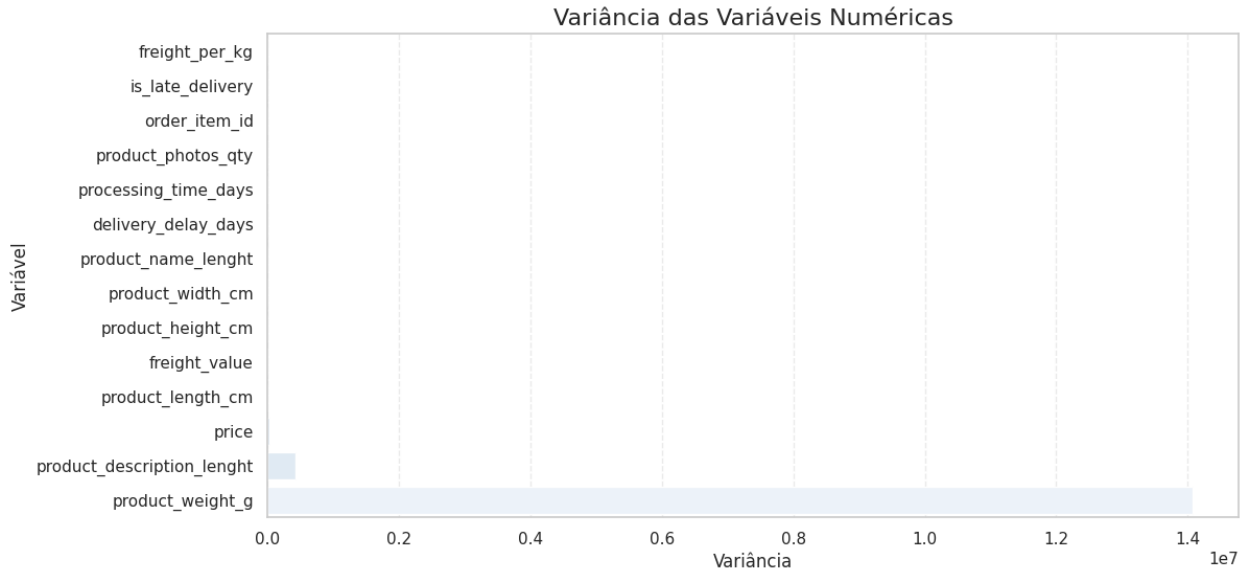
VISUALIZAÇÕES E GRÁFICOS EXPLICATIVOS

A seguir, exibimos as principais visualizações geradas, acompanhadas de interpretações e insights que explicam o que cada gráfico revela. Essas visualizações permitem ver padrões, identificar relações entre variáveis e perceber possíveis problemas ou tendências importantes.

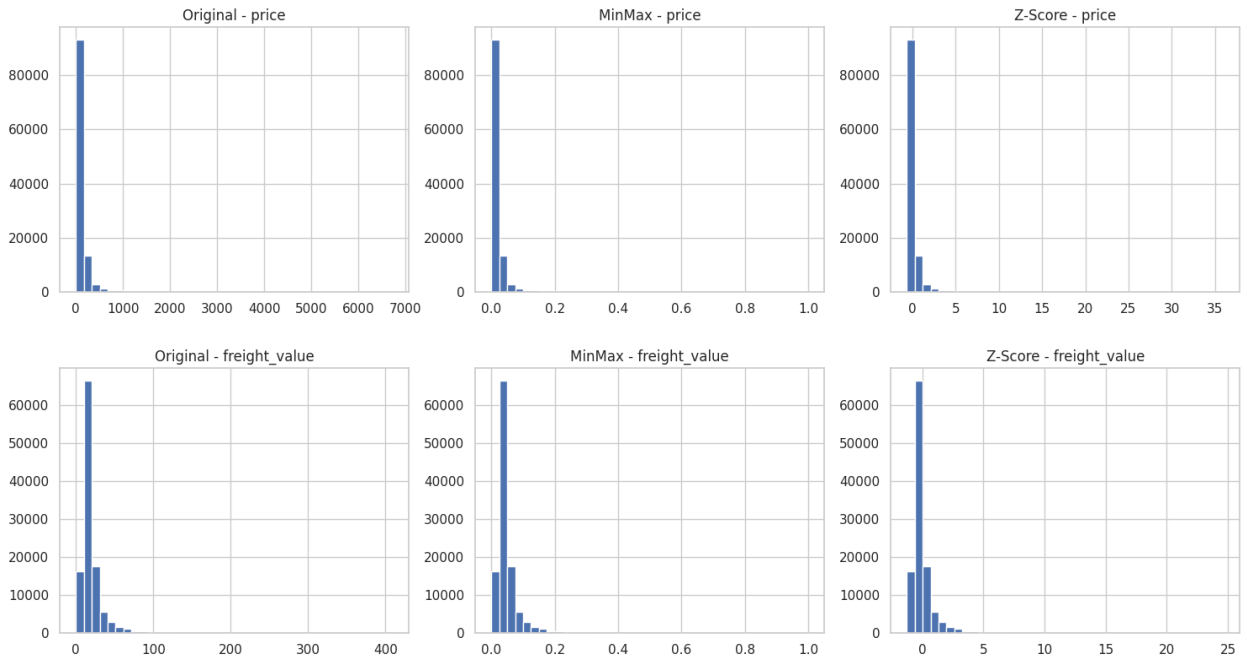
Cada imagem acompanha suas interpretações logo abaixo delas.

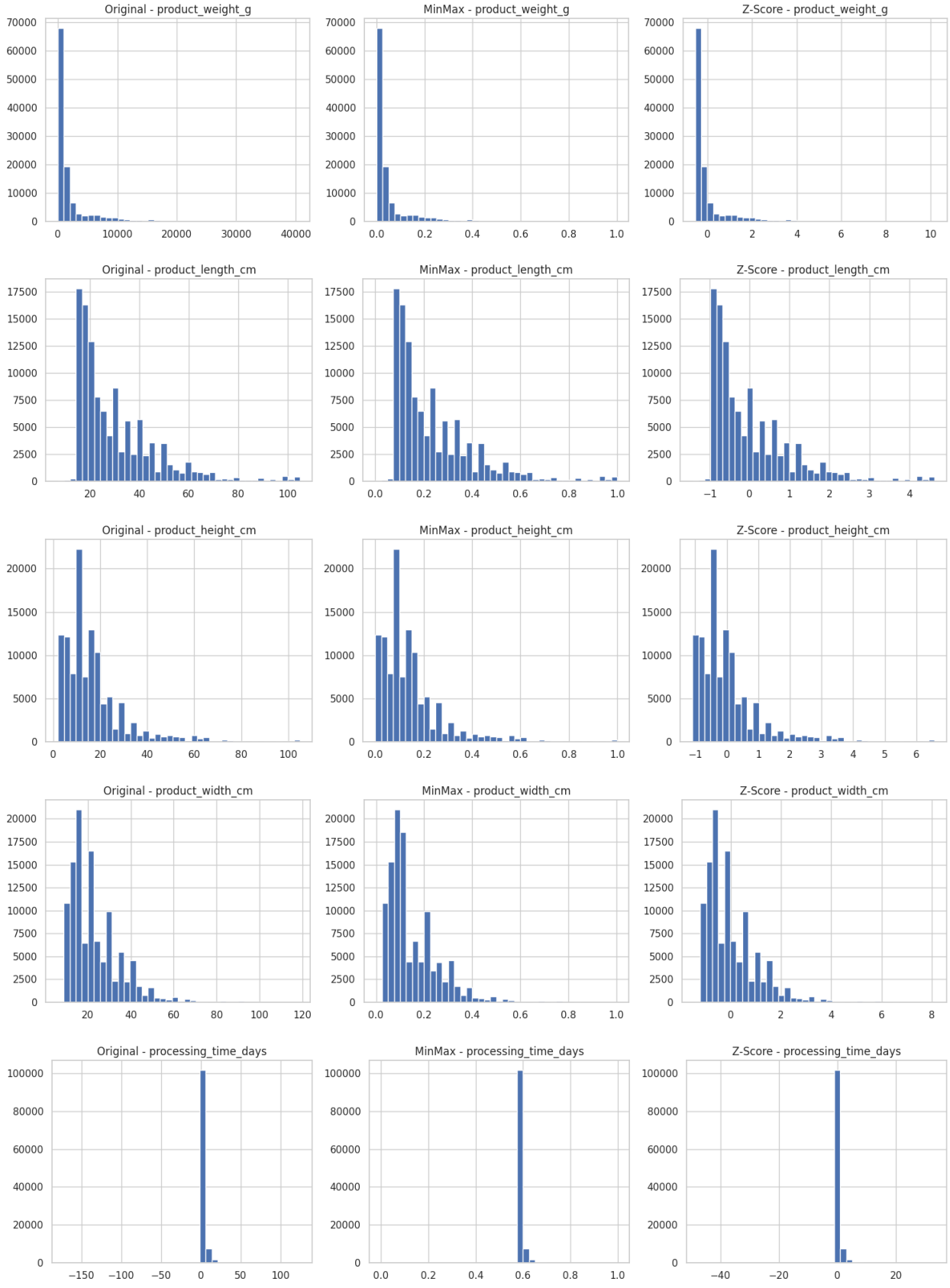


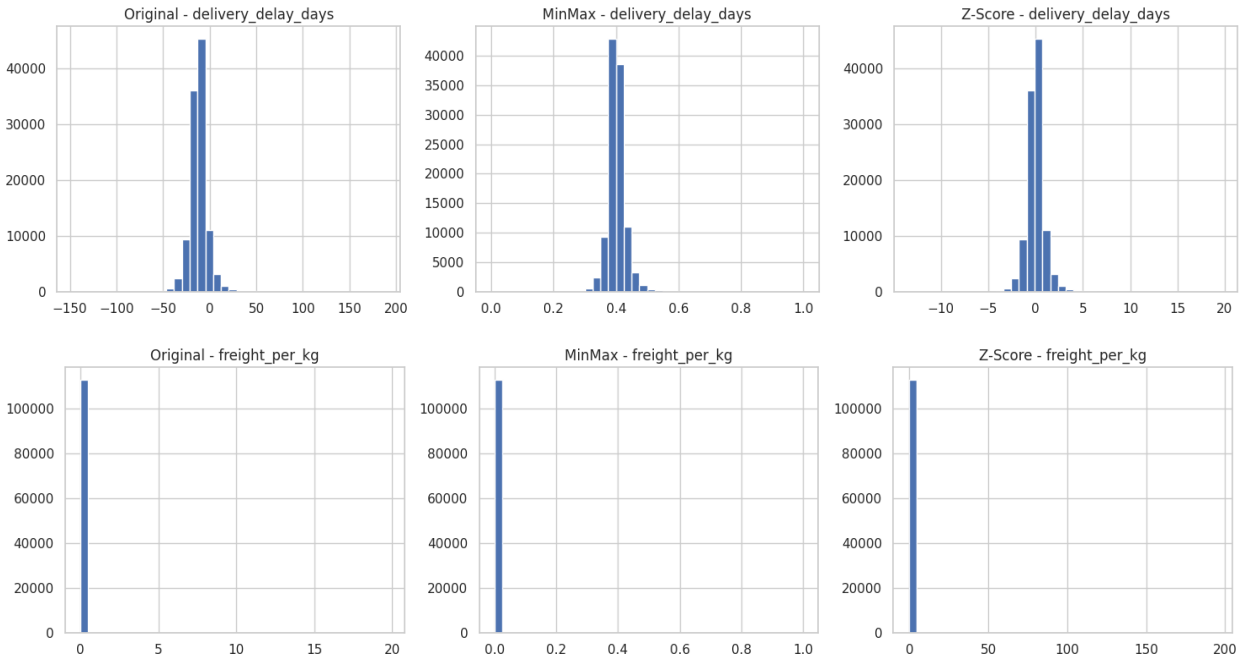
O heatmap mostra expõe que o peso, as dimensões e o frete têm uma forte relação entre eles, ou seja, produtos maiores significam fretes mais caros. Também fica visível que o atraso tem uma relação considerável com tempo de processamento portanto sugere que parte dos atrasos começa com o próprio vendedor.



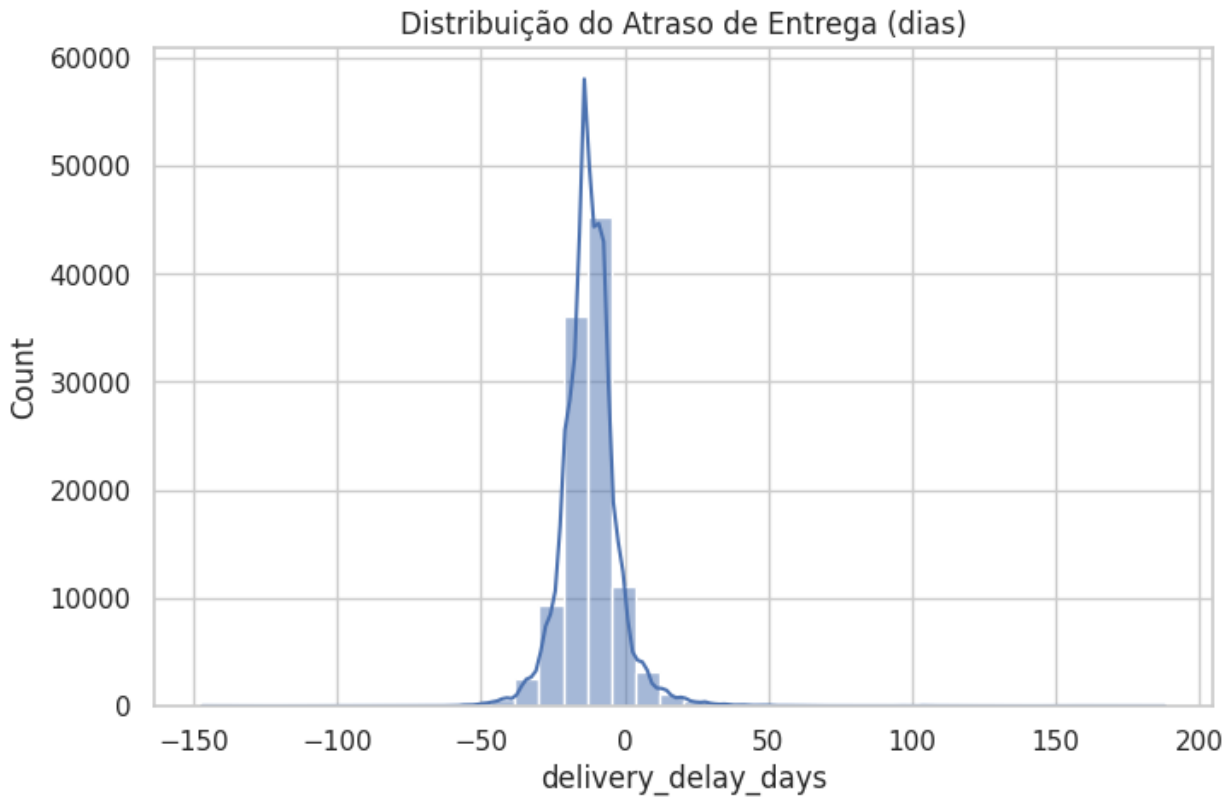
A análise de variância mostra que a coluna "order_item_id" não tem poder explicativo, pois quase não varia. Já as colunas de peso, dimensões e frete apresentam alta variância expondo que impactam na logística.



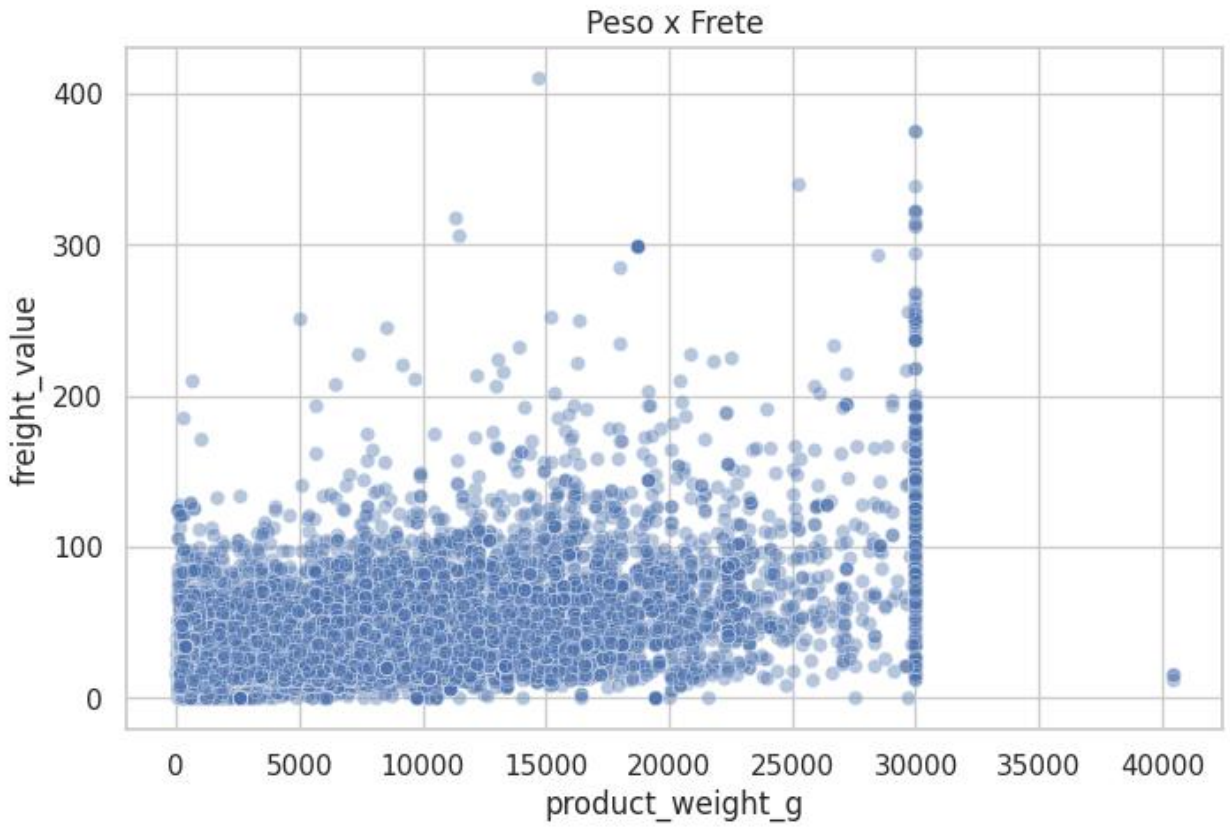

Original - product_weight_g
MinMax - product_weight_g
Z-Score - product_weight_g
Original - product_length_cm
MinMax - product_length_cm
Z-Score - product_length_cm
Original - product_height_cm
MinMax - product_height_cm
Z-Score - product_height_cm
Original - product_width_cm
MinMax - product_width_cm
Z-Score - product_width_cm
Original - processing_time_days
MinMax - processing_time_days
Z-Score - processing_time_days



Os histogramas mostram que mesmo com as transformações a distribuição se mantém, portanto apenas deixa melhor de se comparar atributos e não muda o resultado.

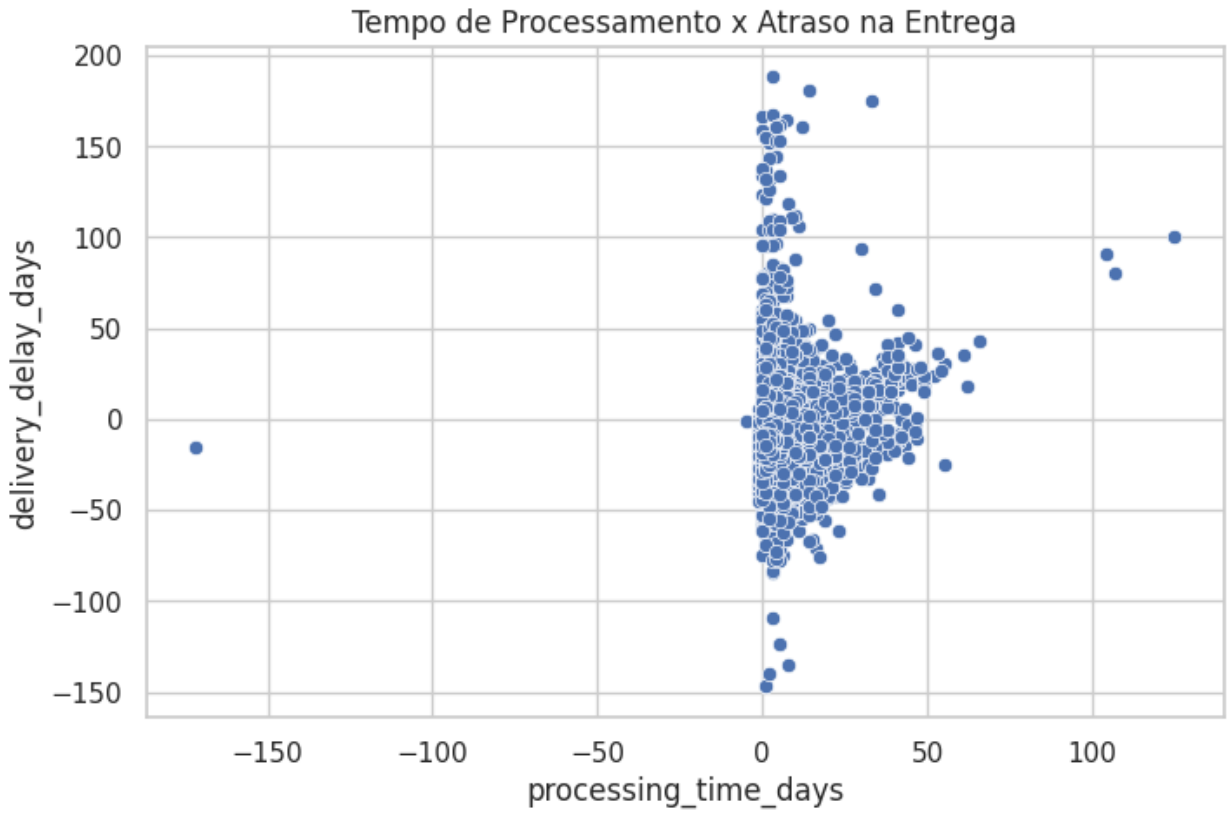


O gráfico mostra uma distribuição altamente concentrada próxima de zero, indicando que a grande maioria dos pedidos é entregue no prazo ou com poucos dias de adiantamento/atraso, isso confirma que a operação logística é eficiente, com pouquíssimos casos de atraso expressivo.



O gráfico mostra uma tendência clara, conforme o peso do produto aumenta, o valor do frete também tende a aumentar. Isso acontece pois geralmente o custo de transporte é diretamente influenciado pelo peso do item.

Também observamos algumas exceções, especialmente em produtos muito pesados com fretes proporcionalmente baixos, ou produtos leves com fretes elevados. Podem ser explicados por fatores como distância da entrega, tipo de produto, política de frete do vendedor ou inconsistências no cadastro do peso.



O gráfico mostra que pedidos com maior tempo de processamento pelo vendedor apresentam maior atraso na entrega final. Essa relação reforça a ideia de que o atraso não ocorre apenas na etapa de transporte, mas também na preparação e separação do pedido. Embora existam valores dispersos, a tendência geral confirma que processamento lento é um forte indicador de possíveis atrasos na entrega.

INSIGHTS FINAIS

Os insights finais foram:

A eficiência logística depende tanto da transportadora quanto do vendedor

Atrasos não são causados apenas pela fase de transporte. O tempo de processamento do pedido pelo vendedor tem forte impacto na entrega final. Vendedores que demoram mais para enviar o pedido tendem a gerar entregas mais atrasadas.

Peso e dimensões são fatores centrais na composição do frete

O frete apresenta correlação clara com o peso dos produtos, reforçando a importância do peso e volume no custo logístico. Produtos pesados ou volumosos naturalmente geram fretes mais elevados.

A grande maioria das entregas é realizada no prazo ou até antes

A distribuição dos atrasos mostra que a operação da Olist é estável, com poucas ocorrências extremas de atraso. Isso indica previsibilidade e boa performance geral do sistema.

Os dados apresentam inconsistências que precisam de atenção

Foram identificados pesos iguais a zero, textos inconsistentes e valores ausentes em massa na tabela de produtos. O pré-processamento foi fundamental para corrigir e padronizar essas informações.

As novas features que ajudaram a explicar o comportamento logístico

A criação das variáveis:

- delivery_delay_days;
- is_late_delivery;
- processing_time_days;
- freight_per_kg.

permitiram enxergar o problema de forma mais estruturada e aprofundada, revelando relações que não estariam visíveis apenas com os dados originais.

Produtos diferem amplamente em características físicas e impacto logístico

Itens mais pesados, com maiores dimensões, e categorias específicas (como móveis) apresentam comportamento distinto no frete e no tempo de entrega, o que pode ajudar a empresa a aprimorar estratégias de precificação e oferta.

Com esses insights também conseguimos responder as perguntas norteadoras pedidas no projeto:

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

Observamos a partir dos gráficos e da feature delivery_delay_days que o principal fator associado ao atraso é o tempo de processamento do vendedor (processing_time_days).

Pedidos em que o vendedor demora mais para entregar o item à transportadora tendem a apresentar maior atraso na entrega final. Isso indica que parte do atraso não ocorre no transporte, mas sim na **preparação do pedido**. Também tem outros fatores com menor influência como produtos com dimensões maiores e produtos de maior peso.

2. Existem categorias de produtos com maior frequência de problemas?

Sim, identificamos que categorias com produtos maiores e mais pesados tendem a apresentar fretes mais altos, maior tempo de transporte e maior risco de atraso. Algumas das categorias são: Móveis, utilidades domésticas grandes, eletrodomésticos, produtos de construção e ferramentas pesadas.

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

Sim, os dados apresentam outliers importantes, por isso decidimos manter os outliers reais e remover/ajustar apenas os inválidos, pois muitos outliares são reais como móveis muito grandes e produtos muito caros. Remover eles distorceriam a distribuição real da plataforma.

Tratamos da seguinte forma:

Pesos iguais a 0 ou negativos foram considerados inválidos e foram substituídos pela mediana da categoria do produto, para manter coerência.

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

Correlação com Frete (freight_value)

Peso do produto (product_weight_g), tem uma correlação forte

Largura, altura e comprimento, tem uma correlação moderada

O frete depende principalmente do peso bruto e das dimensões físicas.

Correlação com Preço (price)

Correlação fraca com praticamente todas as variáveis numéricas

Ou seja, que o preço depende muito mais da categoria do produto do que das características físicas.

Correlação com Atraso (delivery_delay_days)

Correlação moderada com processing_time_days

O atraso é mais influenciado pelo tempo que o vendedor leva para processar o pedido do que pelas dimensões do produto, embora ainda exista influência.

5. A limpeza alterou o comportamento dos dados? Como e por quê?

Sim, a limpeza impactou diretamente nas análises, pois ficou mais coerente principalmente na parte relacionada a frete, atraso e medidas do produto. Melhorou a qualidade analítica sem falsificar os padrões reais dos dados.

- Pesos iguais a zero e valores negativos foram substituídos pela mediana por categoria. Isso corrigiu artefatos de cadastro que distorciam médias e correlações.
- Quando padronizamos os nomes, evitamos que as mesmas categorias aparecessem fragmentadas, o que melhorou cálculos por categoria.
- A conversão de tipos possibilitou cálculo correto como tempo de entrega, o que alterou estatísticas e permitiu criar features relevantes.
- Adotamos a política de não remover outliers legítimos, tratando apenas valores inválidos, o que preservou a variabilidade do dataset, como móveis muito caros.

CONCLUSÕES DO GRUPO

O projeto permitiu aplicar na prática todo o ciclo de vida da Ciência de Dados, desde a compreensão do problema até a preparação final do dataset e entrega. Por meio da exploração dos três principais datastes, foi possível entender o comportamento dos pedidos e extrair informações valiosas sobre o fluxo logístico da plataforma.

A etapa de pré-processamento mostrou-se essencial, envolvendo limpeza de duplicatas, correção de inconsistências, tratamento de valores ausentes, conversão de tipos, padronização de textos, codificação categórica e normalização de variáveis numéricas. Garantiram que os dados estivessem completos, coerentes e prontos para análises mais avançadas.

As visualizações contribuíram significativamente para interpretar padrões importantes, como a relação entre peso e frete, a influência do tempo de processamento no atraso da entrega e análise de envios pontuais. A criação de quatro novas features também trouxe profundidade analítica ao dataset, permitindo enxergar relações que não eram possíveis de serem vistas apenas nos dados originais.

Entendemos que o estudo dos datastes reforçou que o desempenho logístico depende de múltiplos fatores que vão de características físicas dos produtos até a eficiência operacional dos vendedores. O dataset final construído serve como base sólida para análises mais complexas, modelagem preditiva ou até propostas de otimização logística.

Assim, o trabalho cumpre o objetivo de demonstrar o ciclo completo de preparação e análise de dados, seguindo boas práticas utilizadas no mercado e aprendidas em sala.