UNIVERSIDADE PRESBITERIANA MACKENZIE

TECNÓLOGO EM BANCO DE DADOS

Ryan Rodrigues Pereira – 10742607 – 10742607@mackenzista.com.br

Nour Hussein Barakat – 10738273 – 10738273@mackenzista.com.br

Guilherme de Araújo Esp. Santo – 10746294 – 10746294@mackenzista.com.br

ANÁLISE EXPLORATÓRIA DE DADOS

SÃO PAULO

2025

Sumário

1 Contexto 3

1.1 Sobre a Empresa 3

1.2 Problema da Pesquisa 4

1.3 Objetivo e metas 5

2 Cronograma 6

2.1 Funções 6

2.2 Pensamento Computacional 7

3 Dataset 7

3.1 Aquisição 7

3.2 Descrição origem 7

3.3 Descrição dataset 8

4 Metadados e análise exploratória 8

4.1 Metadados dos Principais Datasets: 8

4.2 Limpeza dos Dados (Data Cleaning) 11

4.3 Analise Exploratoria dos Dados: 12

5 Proposta analítica 15

6 Data Storytelling 21

7 Conclusão 22

8 Glossário 22

9 Referências 22

10 Sumário 22

1 Contexto

O Olist é um dos maiores marketplaces brasileiros, conectando pequenos vendedores a clientes em todo o país através de grandes plataformas de e-commerce. Nesse ambiente competitivo, compreender os fatores que influenciam a experiência do consumidor é essencial para fortalecer a confiança dos clientes, otimizar operações logísticas e aumentar a performance dos vendedores.

As avaliações (review scores) desempenham um papel central nesse processo, pois refletem diretamente a percepção do cliente sobre o atendimento, a entrega e a qualidade do produto. Entretanto, múltiplos fatores podem influenciar essas notas, incluindo atrasos na entrega, localização geográfica, categoria do produto e desempenho dos vendedores.

Este projeto utiliza dados públicos do Olist dataset (Kaggle), que reúne informações detalhadas sobre pedidos, entregas, produtos, vendedores e avaliações de clientes. A partir dessa base, buscaremos explorar quantitativamente como esses fatores impactam a satisfação do consumidor e quais padrões podem ser extraídos para orientar estratégias de melhoria.

1.1 Sobre a Empresa

A Olist é uma empresa brasileira de tecnologia para o varejo que oferece soluções para lojas online e offline venderem mais. Ela atua como um ecossistema que conecta vendedores a canais de venda, como grandes marketplaces (Americanas, Mercado Livre, entre outros), e simplifica a gestão de negócios. As soluções incluem ferramentas para centralizar o gerenciamento de produtos, vendas e estoque, além de logística para entrega.

Fundada em fevereiro de 2015 por Tiago Dalvi, a Olist serviu como uma derivação da Solidarium, empresa criada em 2007 como uma loja de shopping que depois se tornou um marketplace para a venda de produtos artesanais. A mudança no modelo de negócio ocorreu após uma aceleração e rodada de

investimento realizada pelo fundo 500 Startups. Outros fundos como Redpoint eventures, Valor Capital Group e SoftBank também são investidores da empresa.

-Missão, visão e valores:

A missão da Olist é empoderar lojistas de todos os portes a venderem mais e melhor online, conectando-os a grandes marketplaces e oferecendo soluções integradas de logística, atendimento e gestão. A empresa valoriza transparência, inovação e crescimento colaborativo.

- Segmento de atuação:

A Olist atua no setor de e-commerce, oferecendo uma plataforma de integração que conecta pequenos e médios varejistas a grandes marketplaces, como Mercado Livre, Amazon e Magalu.

-Market Share / importância no mercado:

A Olist é uma das principais plataformas de integração de e-commerce do Brasil, com presença em diversos marketplaces e forte relevância no ecossistema de varejo digital. Seu papel é estratégico para o setor, pois facilita o acesso de pequenos lojistas a um público nacional, promovendo inclusão digital e competitividade.

-Número de colaboradores:

Atualmente, a empresa possui cerca de 700 colaboradores e mantém escritórios em Curitiba, São Paulo e Bento Gonçalves, atendendo lojistas e marcas de diversos setores do varejo. [1]

1.2 Problema da Pesquisa

Apesar do crescimento do e-commerce no Brasil, muitos clientes ainda demonstram insatisfação nas avaliações de seus pedidos. Entretanto, não está claro quais fatores mais influenciam essas notas; seriam os atrasos na entrega, as diferenças regionais na logística, ou as características dos produtos?

Assim, o problema central que orienta esta pesquisa é:

Quais fatores mais impactam as notas de avaliação dos clientes na plataforma Olist ? E até que ponto os atrasos são determinantes nessa percepção de satisfação?

1.3 Objetivo e metas

O objetivo principal deste estudo é entender como a localização geográfica, a categoria do produto e os atrasos nas entregas impactam as notas de avaliação (review scores) dos clientes

Para isso, pretendemos:

1. Analisar a relação entre atrasos de entrega e notas de avaliação:
   - *Hipótese 1*: Pedidos com atrasos significativos recebem notas médias mais baixas.

2. Identificar categorias de produtos com maior incidência de avaliações negativas:
   - *Hipótese 2:* Categorias com prazos de entrega mais críticos (ex: eletrônicos, presentes) mostram maior sensibilidade a bad reviews.

3. Explorar diferenças regionais na satisfação do cliente:
   - *Hipótese 3:* Regiões com infraestrutura logística menos desenvolvida apresentam maiores atrasos e menores avaliações.

4. Verificar se o atraso é o único fator determinante ou se existem outros aspectos relevantes que reduzem as notas.

2 Cronograma

Mês	Fase	Resultados
Setembro	Preparação e Exploração dos Dados	- Lista de datasets carregados. - Documento com primeiras impressões e notas sobre a qualidade dos dados (ex: % de nulos em cada coluna crítica).
Outubro	Limpeza e Transformação	- DataFrame master limpo e consolidado. - Código de transformação documentado.
Outubro-Novembro	Análise e Visualização	- Conjunto de visualizações e estatísticas resumidas. - Insights preliminares documentados.
Novembro	Consolidação e Apresentação	- Apresentação final com a narrativa dos dados. - Relatório técnico completo.

2.1 Funções

Integrante	Função no Projeto
Ryan Rodrigues Pereira	Responsável pela limpeza dos dados, análise exploratória e elaboração dos gráficos.
Nour Hussein Barakat	Responsável pela documentação do projeto,metadados e análise exploratória, e proposta analítica.
Guilherme de Araújo Esp. Santo	Responsável pela revisão textual, estruturação do relatório, storytelling e integração dos datasets.

2.2 Pensamento Computacional

No projeto com o dataset da Olist, aplicamos o **pensamento computacional** dividindo o problema em etapas (**coleta dos dados, limpeza, cálculo do atraso e análise por vendedor/região/categoria**), identificando padrões nos dados históricos, abstraindo apenas as variáveis mais relevantes (datas de entrega, estados, categorias e notas de avaliação) e criando algoritmos simples em Python/Pandas para calcular métricas, agrupar informações e gerar visualizações que mostram como cada fator impacta as avaliações dos clientes.

3 Dataset

Este capítulo descreve a origem, aquisição e estrutura do conjunto de dados utilizado para conduzir a análise proposta neste projeto.

Repositório de Código: O código completo para análise e visualizações está disponível em:

https://github.com/oR1an/EDA_Brazilian_E-Commerce

3.1 Aquisição

O dataset foi adquirido por meio da plataforma Kaggle, um repositório online de conjuntos de dados para ciência de dados e aprendizado de máquina. O dataset específico intitulado ' Brazilian E-Commerce Public Dataset by Olist' foi fornecido pela Olist, a maior loja de departamentos em marketplaces brasileiros e pode ser acessado no seguinte endereço:

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data?select=olist_products_dataset.csv

O conjunto de dados tem informações de 100 mil pedidos feitos entre 2016 e 2018 em diversos marketplaces no Brasil.

3.2 Descrição origem

A base de dados utilizada neste projeto é de origem oficial, disponibilizada pela empresa Olist no Kaggle, com a colaboração de Francisco Magioli

(Editor), Leo Dabague (Editor) e André Sionek (Admin). Trata-se de uma fonte pública e amplamente utilizada em pesquisas acadêmicas, refletindo dados reais da plataforma de e-commerce Olist, ainda que eventuais limitações metodológicas possam existir.

3.3 Descrição dataset

Dos muitos conjuntos de dados diferentes fornecidos, estamos interessados apenas em quatro conjuntos de dados que usaremos para a análise:

1. olist_orders_dataset: um conjunto de dados sobre os pedidos dos clientes.
2. olist_product_dataset: um conjunto de dados sobre os produtos.
3. olist_order-reviews_dataset: um conjunto de dados sobre as avaliações.
4. Olist_geolocation_dataset: um conjunto de dados com informações sobre as geolocalizações.

4 Metadados e análise exploratória

4.1 Metadados dos Principais Datasets:

1. Dataset: olist_orders_dataset.csv:
   Linhas: 99441, Colunas: 8
   Identificando valores nulos:

- order_id 0
- customer_id 0
- order_status 0
- order_purchase_timestamp 0
- order_approved_at 160
- order_delivered_carrier_date 1783
- order_delivered_customer_date 2965
- order_estimated_delivery_date 0

Variavel	Descrição	Tipo
order_id	identificador exclusivo do pedido	Int

Order_item_id	número sequencial que identifica o número de itens incluídos na mesma ordem	Int
Product_id	Identificador exclusivo do produto	Int
Seller_id	Identificador exclusivo do vendedor	Int
Shipping_limit_date	Exibe a data limite de envio do vendedor para processar o pedido ao parceiro logístico.	Date
Price	preço do item	Double
Freight_value	valor do frete do item (se um pedido tiver mais de um item, o valor do frete será dividido entre os itens)	Double

2. Dataset: olist_product_dataset.csv:
   Linhas: 32951, Colunas: 9
   Identificando valores nulos:

- product_id 0
- product_category_name 610
- product_name_lenght 610
- product_description_lenght 610
- product_photos_qty 610
- product_weight_g 2
- product_length_cm 2
- product_height_cm 2
- product_width_cm 2

Variavel	Descrição	Tipo
product_id	identificador exclusivo do produto	Int
product_category_name	categoria do produto, em português.	String
product_name_lenght	número de caracteres extraídos do nome do produto.	Int
product_description_lenght	número de caracteres extraídos da descrição do produto.	Int
product_photos_qty	número de fotos publicadas do produto	Int
product_weight_g	peso do produto medido em gramas.	Double

product_length_cm	comprimento do produto medido em centímetros.	Double
product_height_cm	altura do produto medida em centímetros.	Double
product_width_cm	largura do produto medida em centímetros.	Double

3. Olist_order_reviews_dataset.csv:
   Linhas: 99224, Colunas: 7
   Identificando valores nulos:

- review_id 0
- order_id 0
- review_score 0
- review_comment_title 87656
- review_comment_message 58247
- review_creation_date 0
- review_answer_timestamp 0

Variavel	Descrição	Tipo
review_id	identificador exclusivo da avaliação	Int
order_id	identificador exclusivo do pedido	Int
review_score	Nota de 1 a 5 dada pelo cliente em uma pesquisa de satisfação.	Int
review_comment_title	Título do comentário da avaliação deixada pelo cliente, em português.	String
Review_comment_message	Mensagem do comentario da avaliação deixada pelo cliente, em português.	String
Review_creation_date	Mostra a data em que a pesquisa de satisfação foi enviada ao cliente.	Date
review_answer_timestamp	Mostra o carimbo de data/hora da resposta da pesquisa de satisfação.	Time

4. Olist_geolocation_dataset.csv:
   Linhas: 99441, Colunas: 5
   Identificando valores nulos:

- geolocation_zip_code_prefix 0

- geolocation_lat 0
- geolocation_lng 0
- geolocation_city 0
- geolocation_state 0

Variavel	Descrição	Tipo
geolocation_zip_code_prefix	5 primeiros dígitos do CEP	Int
Geolocation_lat	latitude	Double
Geolocation_lng	longitude	Double
geolocation_city	nome da cidade	String
geolocation_state	Nome do estado	String

4.2 Limpeza dos Dados (Data Cleaning)

Durante a etapa de limpeza dos dados, foram aplicadas diversas transformações para garantir a consistência, integridade e relevância das informações utilizadas na análise. As principais ações realizadas foram:

1. Tradução dos atributos para o português:
   Para facilitar a compreensão e a manipulação dos dados, optamos por renomear todos os atributos para o português.
   Exemplo:

categoria_produto	tam_nome_produto	tam_descricao_produto	qtd_fotos_produto	peso_g	comprimento_cm	altura_cm	largura_cm
perfumaria	40.0	287.0	1.0	225.0	16.0	10.0	14.0
artes	44.0	276.0	1.0	1000.0	30.0	18.0	20.0
esporte_lazer	46.0	250.0	1.0	154.0	18.0	9.0	15.0
bebes	27.0	261.0	1.0	371.0	26.0	4.0	26.0
utilidades_domesticas	37.0	402.0	4.0	625.0	20.0	17.0	13.0
...	...	...	...	...	...	...	...
moveis_decoracao	45.0	67.0	2.0	12300.0	40.0	40.0	40.0
erramentas_iluminacao	41.0	971.0	1.0	1700.0	16.0	19.0	16.0
cama_mesa_banho	50.0	799.0	1.0	1400.0	27.0	7.0	27.0
informatica_acessorios	60.0	156.0	2.0	700.0	31.0	13.0	20.0
cama_mesa_banho	58.0	309.0	1.0	2083.0	12.0	2.0	7.0

ıns

2. Tratamento de valores nulos:

- Foram identificados valores ausentes em colunas como order_delivered_customer_date, product_category_name, product_weight_g e review_comment_message.
- Linhas com ausência de datas de entrega foram removidas, pois não é possível calcular o atraso sem essa informação.
- Nos demais casos (como descrições de produto ou mensagens de review ausentes), os valores foram mantidos, uma vez que não comprometem a análise principal.

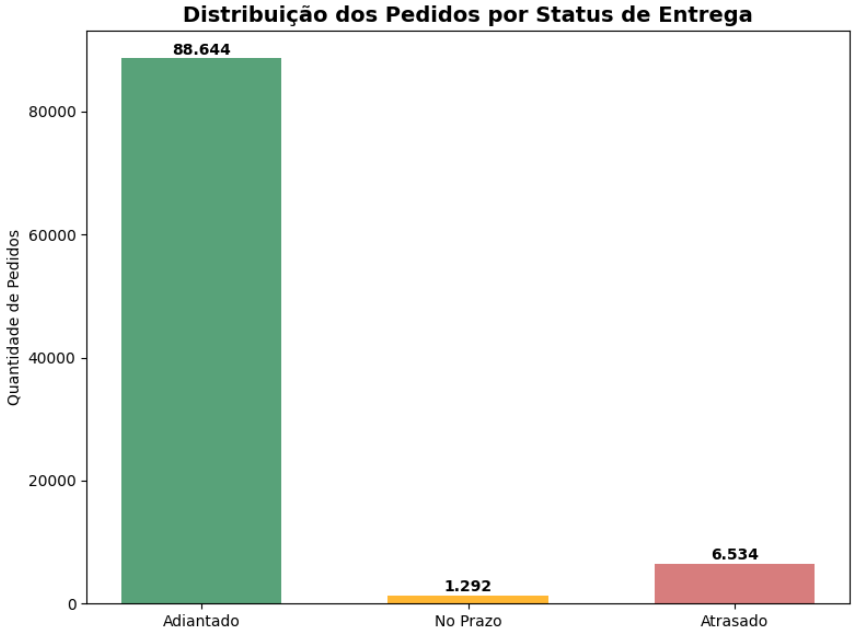
3. Conversão de tipos de dados:
   - As colunas relacionadas a datas (order_purchase_timestamp, order_delivered_customer_date, etc.) foram convertidas para o tipo datetime.
   - Valores numéricos foram ajustados para os tipos adequados (float e int) a fim de permitir cálculos e estatísticas sem erro.
4. Criação de variáveis derivadas:
   Criou-se a variável atraso, calculada pela diferença entre a data_entrega e a data_prevista.
   - Valores positivos indicam dias de atraso.
   - Valores negativos indicam dias adiantados.
5. Filtragem de registros:
   - Mantiveram-se apenas os pedidos com order_status = "delivered", pois são os únicos que possuem data real de entrega.
   - Removeram-se registros duplicados (caso houvesse) e linhas inconsistentes com informações incompletas.
6. Junção de datasets:
   - Os datasets foram unidos utilizando a coluna order_id, conectando informações de pedidos, produtos, avaliações e localização.
   - Essa junção permitiu uma visão integrada do comportamento de entrega e satisfação do cliente.

   Após a limpeza, os dados ficaram prontos para as análises exploratórias e estatísticas, garantindo que os resultados fossem baseados em informações consistentes e confiáveis.

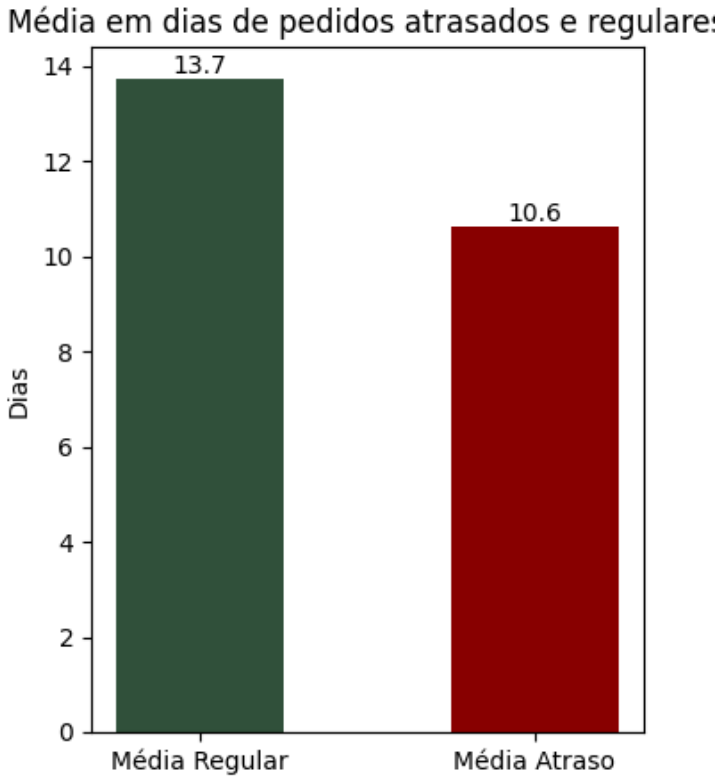
4.3 Analise Exploratoria dos Dados:

Vamos usar alguns métodos básicos de estatística descritiva para visualizar os dados e entender melhor como eles são distribuídos.

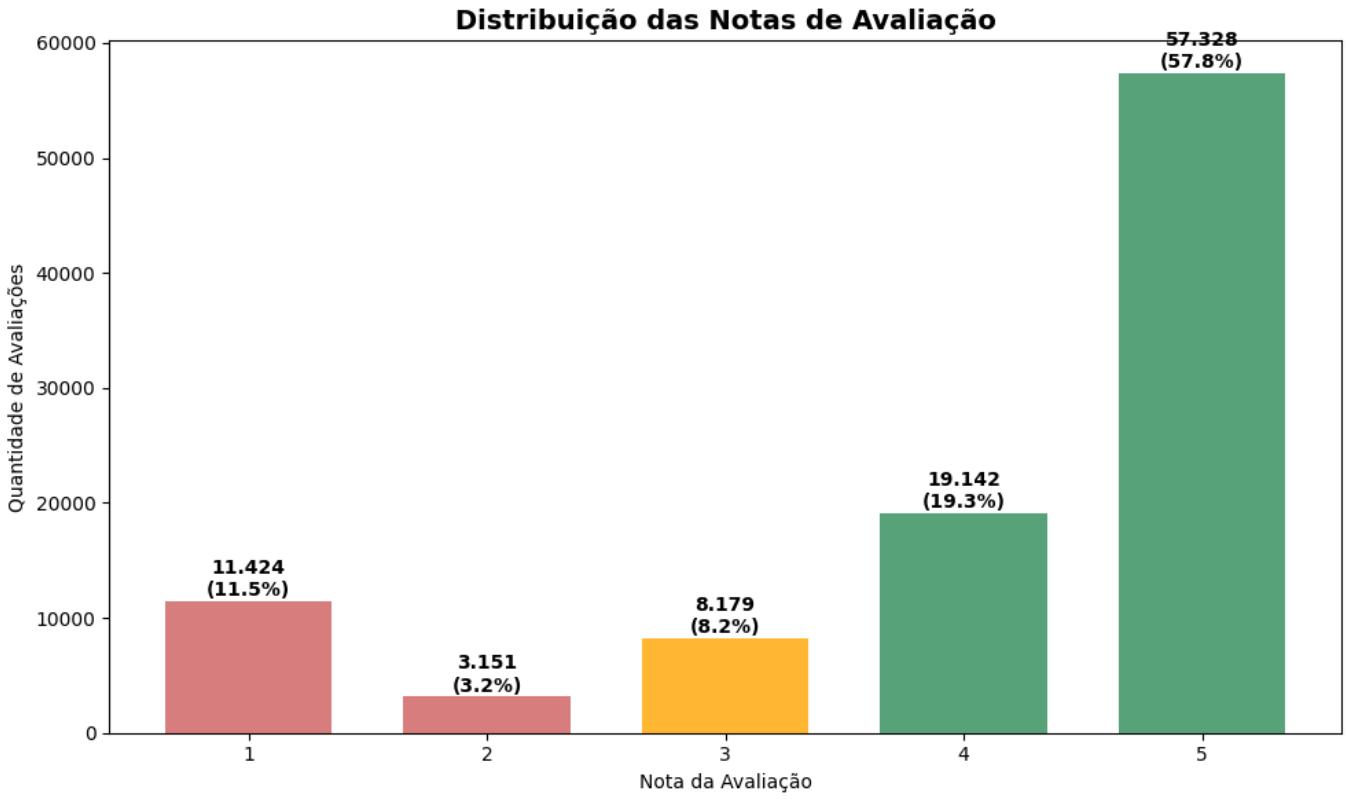
1. **Como as entregas são distribuídas (entregues antes do tempo previsto/no prazo/atrasadas)?**
    - Adiantado 88644
    - Atrasado 6534
    - No Prazo 1292



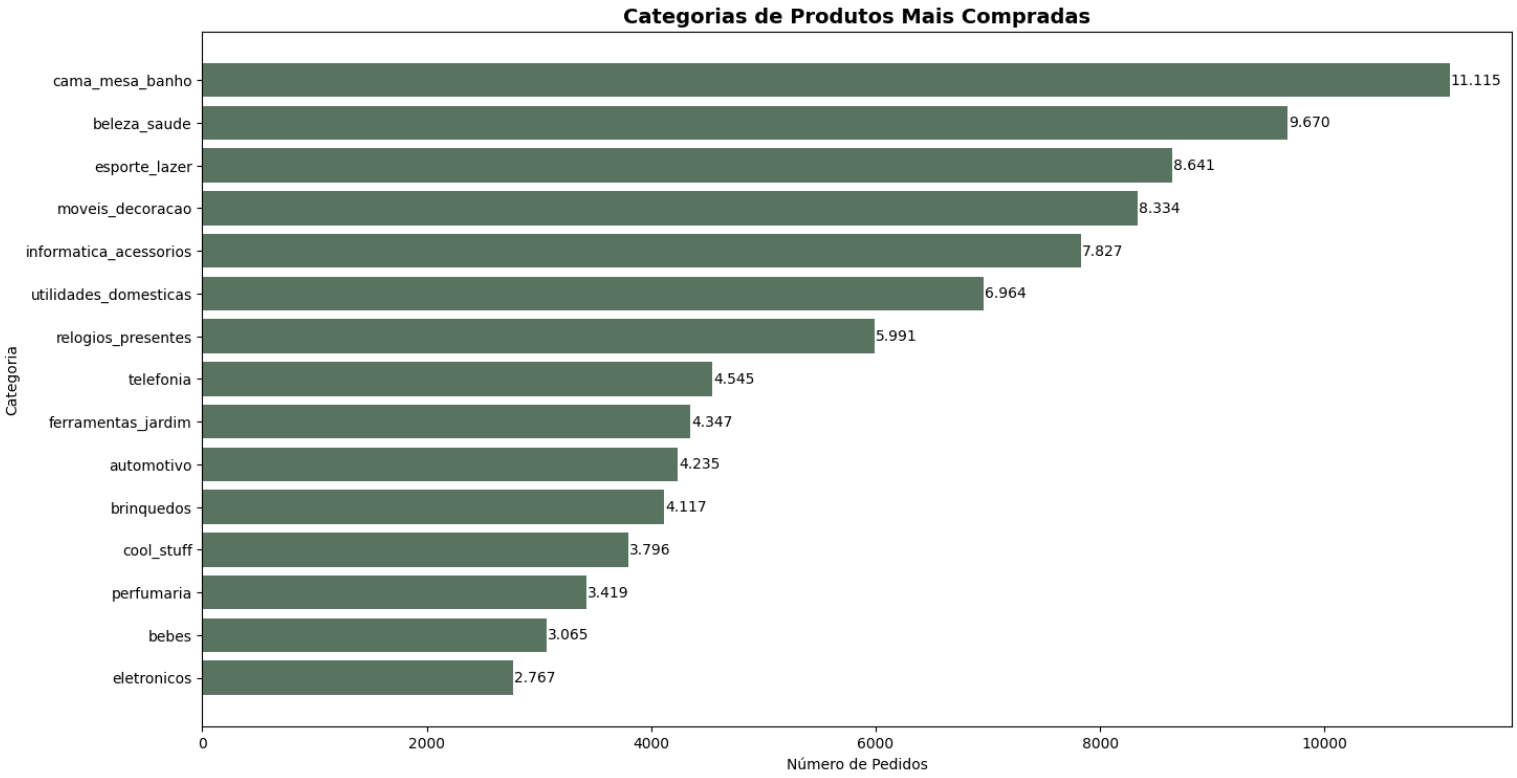
2. Qual é a média em dias de entrega dos pedidos atrasados e regulares? Obs: usamos valores absolutos da media regular para melhor visualização e comparação



3. Como os clientes avaliam suas compras?



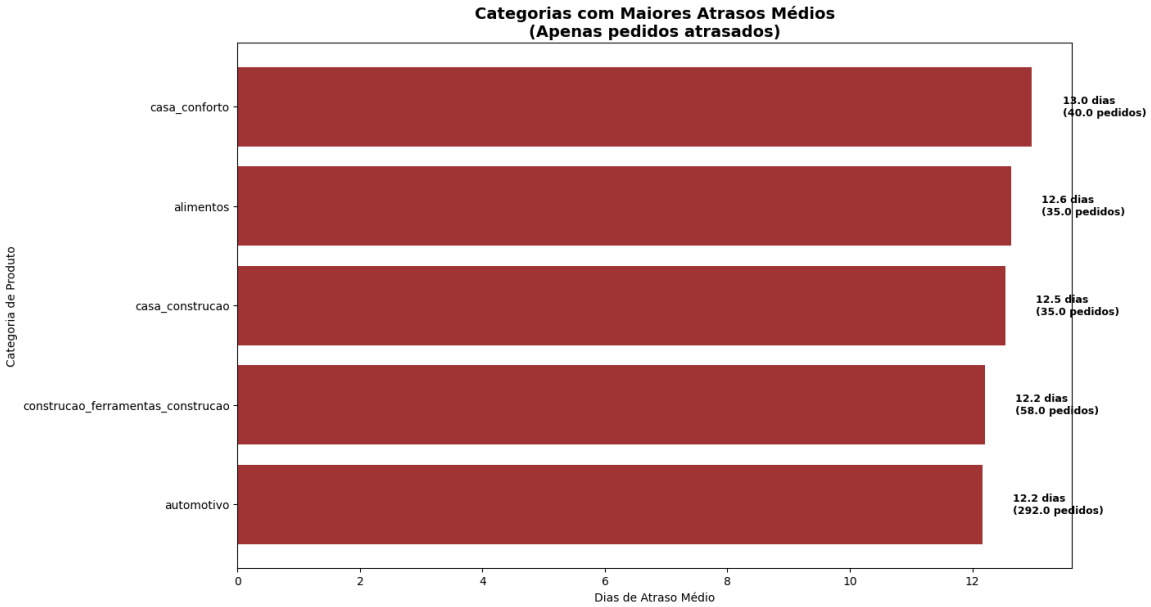
4. Qual é a distribuição de vendas por categoria de produto?



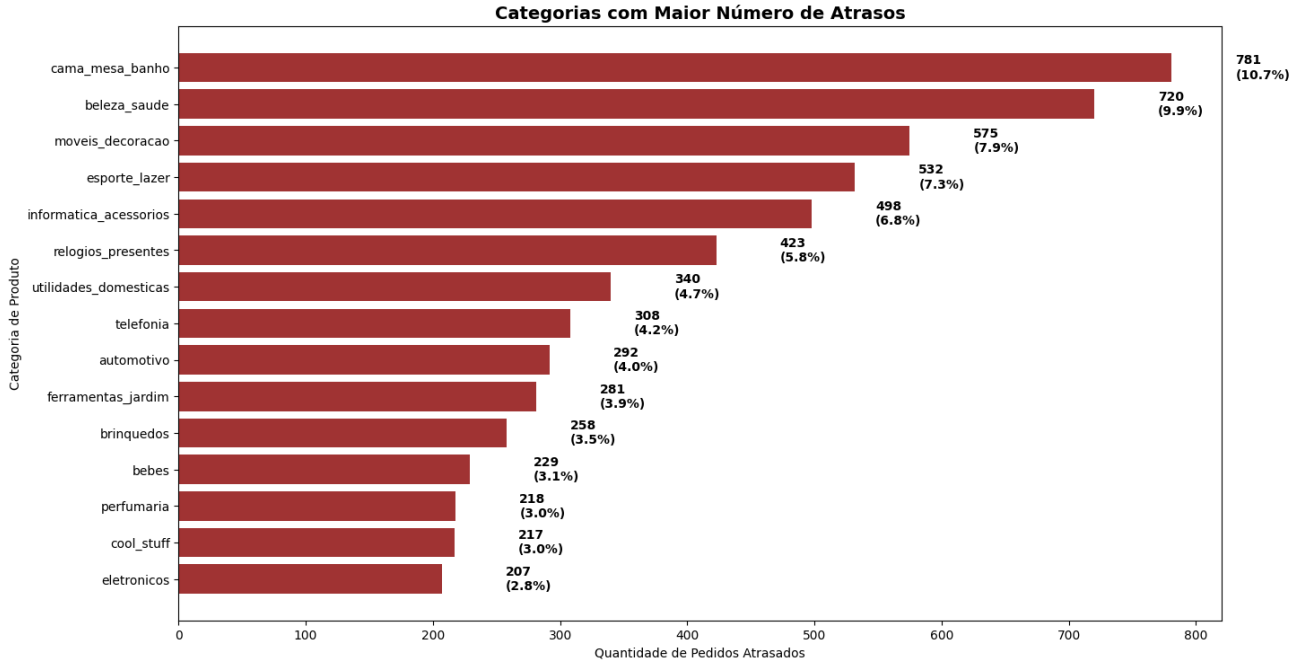
5 Proposta analítica

Tendo mapeado a distribuição dos dados, investigaremos agora como as variáveis se relacionam e que conclusões podemos derivar.

1. Quais produtos costumam atrasar mais?
   Que produtos tem a maior media de atrasos?



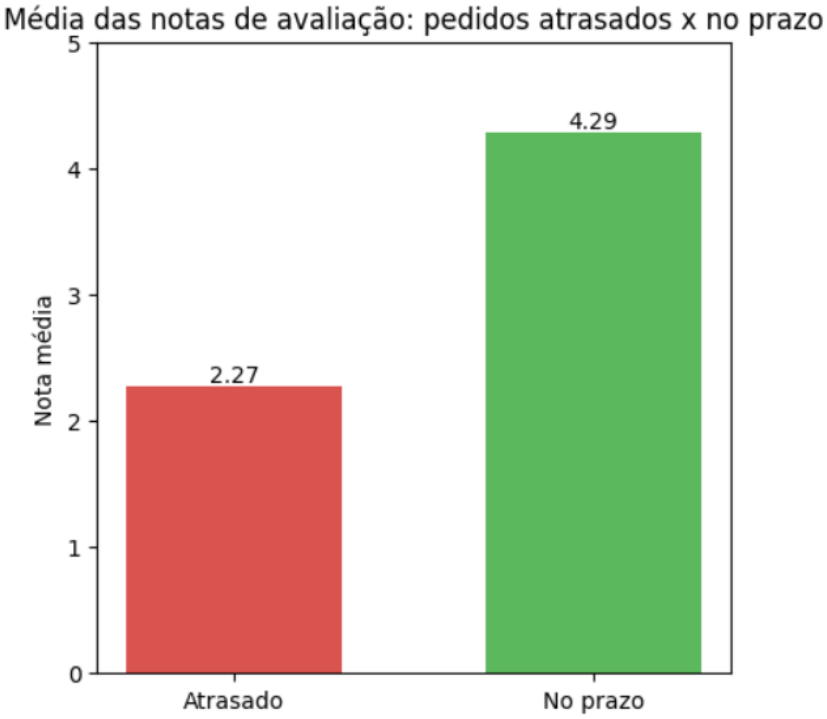
➔ O produto que leva mais dias para ser entregue, é casa_conforto



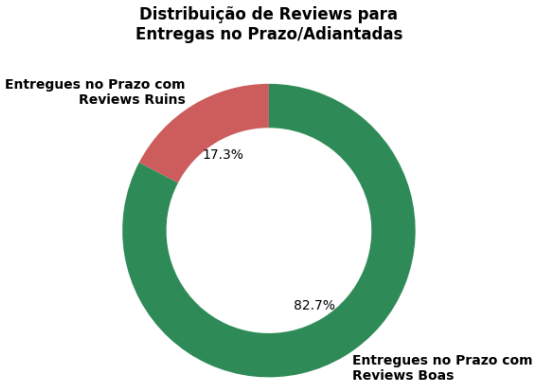
➔ O produto mais atrasado é casa_mesa_banho

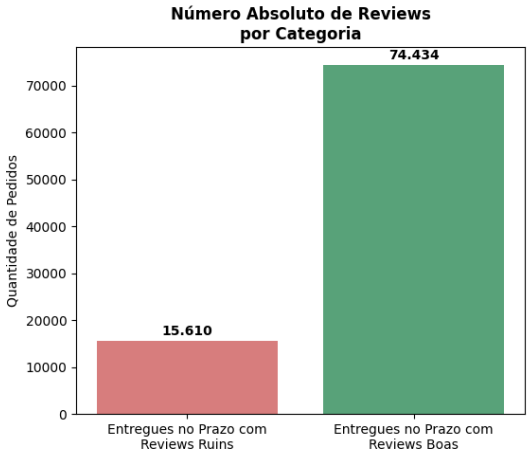
2. Entendendo como pedidos atrasados afetam as avaliações:
   Hipótese 1: Pedidos com atrasos significativos recebem notas médias

mais baixas.

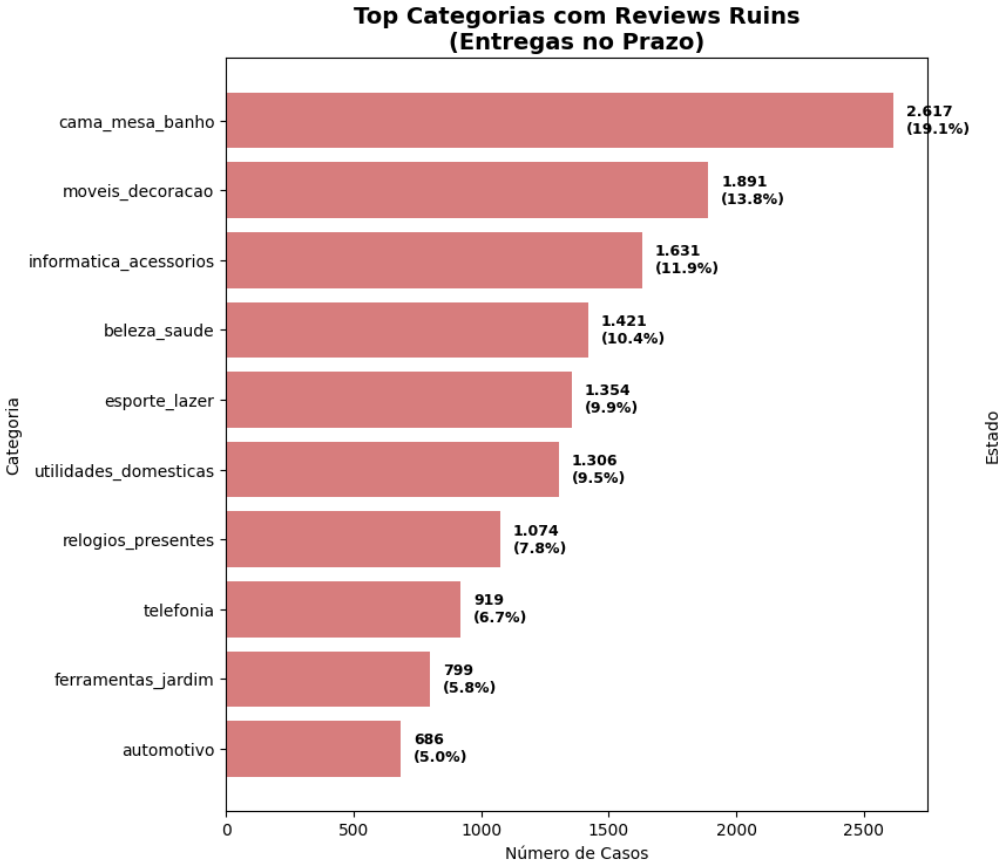


Os dados confirmam a hipotese; atrasos impactam negativamente as avaliações, porém revelam um insight crucial: 17,3% das entregas pontuais ainda geram insatisfação. Isso representa 15.610 pedidos onde fatores além do prazo de entrega influenciaram as avaliações negativas.





Com atrasos respondendo por parte do problema, que outras variáveis - qualidade do produto, comunicação, expectativas - influenciam as avaliações ruins?



Mesmo sendo entregue no prazo, cama_mesa_banho, por exemplo ainda recebe avaliações ruins. Mas por quê? Alguns dos motivos podem ser:
Aqui estão as principais razões pelas quais **cama_mesa_banho** pode receber avaliações ruins mesmo com entregas no prazo:

Problemas Relacionados ao Produto

- Qualidade do material: Roupas de cama que encolhem após lavagem, toalhas que desbotam, tecidos que não correspondem à descrição
-Tamanho impreciso: Lençóis que não servem no colchão, edredons menores que o anunciado
- Cor diferente: Cores que não correspondem às fotos do site (problema comum em e-commerce)

Problemas de Expectativa vs Realidade

- Textura diferente: Tecido áspero quando o cliente esperava maciez
- Espessura inadequada: Edredons muito finos para inverno, cobertores muito quentes
- Acabamento pobre: Costuras mal feitas, bordas desfiando

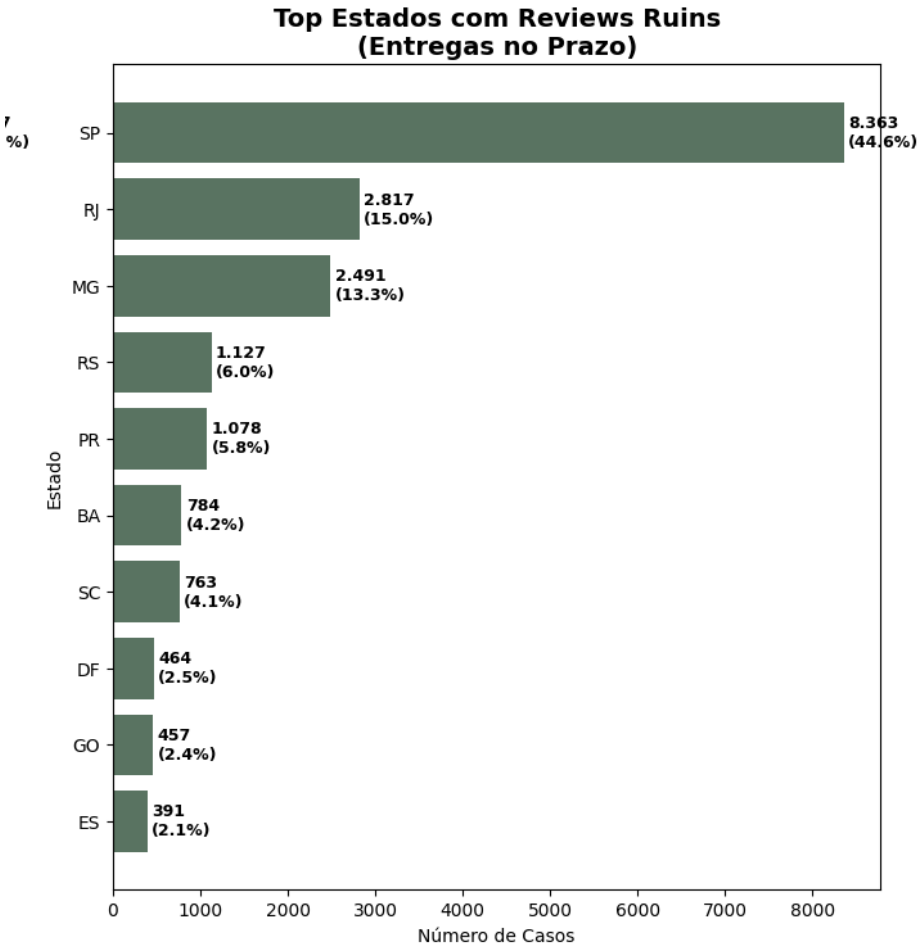
Problemas de Comunicacao
-Descrição incompleta:Não mencionar que o produto precisa de montagem
- Fotos enganosas: Imagens que fazem o produto parecer maior/melhor
- Instruções faltando: Como lavar, medidas exatas

Fatores Específicos da Categoria
- Produtos sensíveis: Roupas de cama e toalhas têm contato direto com a pele
- Altas expectativas: Clientes são mais críticos com produtos que usam diariamente
- Presentes frequentes: Muitas compras são presentes, aumentando a expectativa

Recomendações para Melhoria:
1. Melhorar fotos com escala de referência
2. Descrições detalhadas sobre composição do tecido
3. Incluir vídeos mostrando textura e flexibilidade
4. Amostras de tecido para clientes frequentes
5. Reviews verificadas com fotos reais dos clientes

Esta categoria foi um exemplo de muitos que sofre com o "gap de expectativa" ; o cliente imagina uma experiência de hotel 5 estrelas e recebe um produto comum.
A hipótese *Hipótese 2:* Categorias com prazos de entrega mais críticos (ex: eletrônicos, presentes) mostram maior sensibilidade a bad reviews é desafiada pelos dados. cama_mesa_banho surge como a categoria mais atrasada, porém estes produtos não são tipicamente críticos em termos de tempo. Isso sugere que a pontualidade na entrega sozinha não explica a sensibilidade das avaliações - qualidade do produto, expectativas do cliente e outros fatores desempenham papéis igualmente importantes ou maiores na satisfação do cliente.

3. O que explica a concentração geográfica de avaliações ruins entre pedidos entregues pontualmente?



Apesar do desempenho perfeito de entrega, surgem padrões geográficos significativos na insatisfação dos clientes. A região Sudeste domina as avaliações negativas para entregas pontuais, com São Paulo (40,6%), Rio de Janeiro (13,7%) e Minas Gerais (12,1%) respondendo por quase dois terços de todo o feedback negativo apesar da entrega no prazo. Isso revela que a pontualidade na entrega sozinha não garante satisfação, e fatores regionais desempenham um papel crucial na experiência do cliente.

Fatores Demográficos e de Expectativa:

- SP/RJ/MG: Populações urbanas maiores com expectativas mais elevadas e menor tolerância a imperfeições
- Centros urbanos: Clientes mais experientes em e-commerce e mais críticos em suas avaliações
- Padrões de consumo: Regiões com maior poder aquisitivo podem ter expectativas mais altas de qualidade

Fatores Logísticos e de Infraestrutura:

- Última milha complexa: Entregas em grandes centros urbanos enfrentam mais manuseio, aumentando riscos de danos
- Condições de transporte: Tráfego intenso e más condições de estradas podem afetar a integridade dos produtos
- Centros de distribuição: Possíveis problemas no armazenamento ou manuseio regional

Fatores Culturais e Comportamentais:

- Cultura de avaliação: Clientes no Sudeste podem ser mais propensos a avaliar produtos negativamente
- Familiaridade com tecnologia: Maior comfort com plataformas online para expressar insatisfação
- Expectativas de serviço: Padrões mais altos baseados em experiências anteriores com marketplaces

Fatores Operacionais Específicos:

- Problemas regionais de fornecedores: Vendedores específicos podem estar concentrados nestas regiões
- Comunicação regional: Possíveis gaps na comunicação pós-venda em determinadas áreas
- Problemas de estoque regional: Itens armazenados em centros específicos podem ter issues de qualidade

Insight Estrategico:

A concentração geográfica de insatisfação revela que a empresa precisa de estratégias regionalizadas para melhorar a experiência do cliente, indo além da otimização logística para abordar fatores locais de qualidade, comunicação e gestão de expectativas.

6 Data Storytelling

A história contada pelos dados revela que a pontualidade das entregas é o principal determinante da satisfação.

Produtos de eletrônicos e móveis são os mais suscetíveis a avaliações negativas, enquanto regiões com maior infraestrutura logística tendem a melhores resultados.

Essas descobertas reforçam a necessidade de estratégias regionais de entrega e comunicação proativa com o cliente.

7 Conclusão

A análise quantificou o impacto dos atrasos nas avaliações, mas revelou um insight crucial: categoria do produto e localização são variáveis igualmente determinantes na satisfação do cliente. Estes resultados oferecem métricas acionáveis para otimizar tanto a gestão logística quanto a experiência do consumidor de forma segmentada.

8 Glossário

Atraso - Diferença entre data real e estimada de entrega.

Review Score - Nota de 1 a 5 atribuída pelo cliente.

Categoria - Tipo de produto vendido (ex.: eletrônicos, moda, casa).

UF - Unidade federativa do endereço do cliente.

Seller - Vendedor responsável pelo pedido na plataforma.

9 Referências

1. https://pt.wikipedia.org/wiki/Olist
2. Kaggle – Brazilian E-Commerce Public Dataset by Olist. Disponível em: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
3. Wikipédia – Olist (2024). https://pt.wikipedia.org/wiki/Olist
4. Scielo (2023). Logistics Performance and E-commerce Customer Satisfaction in Brazil.

10 Sumário