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

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 Objetivo e metas

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

Para isso, pretendemos:

- Analisar a relação entre atrasos de entrega e notas de avaliação.

-Identificar vendedores e categorias de produtos com maior incidência de avaliações negativas.

-Explorar diferenças regionais na satisfação do cliente.

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

1.2 Problema da Pesquisa

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

Qual será a função de cada um?

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

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 com as seguintes variáveis:

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. olist_product_dataset: um conjunto de dados sobre os produtos com as seguintes variáveis

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: um conjunto de dados sobre as avaliações com as seguintes variáveis:

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: um conjunto de dados com informações sobre as geolocalizações com as variáveis:

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