

MVP: Sprint: Engenharia de Dados

Clara Delboni

Conjunto de dados: "Brazilian E-Commerce Public Dataset by Olist"

Objetivo

O objetivo deste MVP é analisar o conjunto de dados publicado pela Olist, plataforma intermediadora de marketplace brasileira, com o intuito de responder a perguntas chave relacionadas aos pedidos cancelados. O foco principal será identificar padrões e fatores que influenciam o cancelamento de pedidos, como características dos produtos, dados transacionais e processo de entrega ineficaz. Os insights gerados poderiam reduzir a taxa de cancelamento, maximizando os lucros da empresa e otimizando a experiência do cliente.

Perguntas

Possíveis perguntas que podem ser usadas para explorar insights relacionados ao status de cancelamento do conjunto de dados "Brazilian E-Commerce Public Dataset by Olist" possui:

- Qual a taxa de cancelamento?
- Quais são as categorias com maior número de cancelamento?
- Quais são as categorias com maior taxa de cancelamento?
- Qual a taxa de cancelamento por cidade?
- O tempo de entrega influencia a taxa de cancelamento?
- O tempo de entrega ao transportador influencia a taxa de cancelamento?
- Há clientes que cancelam antes do produto ser entregue?
- A taxa de cancelamento varia dependendo da faixa de preço dos produtos?
- Existe uma correlação entre o tipo de pagamento e o cancelamento de pedidos?
- Há casos em que os pedidos são cancelados após a entrega? Qual o motivo e a frequência disso?
- Qual a taxa de cancelamento de pedidos por cliente, e há clientes com múltiplos cancelamentos?
- Há relação entre descontos e cancelamentos?

Modelagem

No total, possuímos oito tabelas, e embora o modelo em Floco de Neve seja considerado mais complexo para interpretação, ele foi adotado na camada Silver por se adequar melhor à estrutura dos dados deste projeto. Nesse modelo, as tabelas de dimensão são normalizadas e divididas em subdimensões, fazendo sentido com as fontes utilizadas. Um exemplo é a dimensão originada da tabela `olist_order_items_dataset`, que se relaciona com subdimensões como

olist_products_dataset, olist_sellers_dataset e olist_geolocation_dataset, justificando assim a escolha por esse modelo.

No entanto, para fins analíticos e visando otimizar a performance de leitura e facilitar o consumo dos dados, foi criada na camada Gold uma tabela no formato flat, por meio de *joins* selecionados entre colunas de algumas tabelas da Silver. Essa estrutura simplificada permite responder de forma mais ágil às perguntas definidas no início do projeto. Caso seja necessário realizar análises mais detalhadas ou com outra granularidade no futuro, ainda é possível recorrer às tabelas da camada Silver, estruturadas em modelo Floco de Neve.

Aplicação:

- Tabela Fato:
 - olist_orders_dataset – Contém informações sobre o pedido, como id do pedido, id do cliente e outros detalhes sobre cada pedido).
- Tabelas de Dimensão:
 - Itens (olist_order_items_dataset).
 - Clientes (olist_customers_dataset).
 - Pagamentos (olist_order_payments_dataset).
 - Avaliações (olist_order_reviews_dataset).
- Subdimensões (essas tabelas são relacionadas à tabela de itens do pedido):
 - Produtos (olist_products_dataset).
 - Vendedores (olist_sellers_dataset).
 - Geolocalização (olist_geolocation_dataset).

Linhagem dos dados

Os dados foram coletados do conjunto de dados público 'Brazilian E-Commerce Public Dataset by Olist' disponível na plataforma Kaggle. É disposto nas informações sobre o dataset que o conjunto de dados de e-commerce Brasileiro apresentado foi disponibilizado pela empresa Olist e contém informações de aproximadamente 100 mil pedidos de 2016 a 2018 feitos em vários marketplaces no Brasil. Além disso, é informado que foi adicionado um conjunto de dados de geolocalização, que relaciona os CEPs brasileiros às coordenadas de latitude e longitude.

Catálogo de dados – Tabelas camada Silver:

1. Tabela fato - Silver.fatopedidos_final:

Descrição: Contém informações sobre o pedido, como ID do pedido, ID do cliente e outros detalhes sobre cada um.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
Order_id	String	Id do pedido	N/A	N/A	N/A
Customer_id	String	Id cliente	N/A	N/A	N/A
Order_status	String	Status pedido	N/A	N/A	'Entregue', 'Cancelado', 'Processando', 'A caminho'
Order_purchase_timestamp	timestamp	Data e hora da compra	N/A	N/A	N/A
Order_approved_at	timestamp	Data da aprovação do pagamento	N/A	N/A	N/A
Order_delivered_carrier_date	timestamp	Data em que o pedido foi entregue ao transportador	N/A	N/A	N/A
Order_delivered_customer_date	timestamp	Data de entrega	N/A	N/A	N/A
Order_estimated_delivery_date	timestamp	Data estimada de entrega	N/A	N/A	N/A

2. Silver.dimclientes:

Descrição: Contém informações sobre os clientes.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
customer_id	String	Id do cliente em cada pedido	N/A	N/A	
customer_unique_id	Inteiro	Id único do cliente	N/A	N/A	
customer_zip_code_prefix	String	Prefixo CEP cliente	N/A	N/A	
customer_city	String	Cidade do cliente	N/A	N/A	Cidades do Brasil
customer_state	String	Estado do cliente	N/A	N/A	Estados do Brasil

3. Silver.dimlocalizacao:

Descrição: Contém informações sobre a geolocalização, associando CEPs às coordenadas geográficas.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
geolocation_zip_code_prefix	Inteiro	Prefixo do CEP cliente	N/A	N/A	N/A
geolocation_longitude	Double	Longitude do local do CEP	N/A	N/A	N/A
geolocation_latitude	Double	Latitude do local do CEP	N/A	N/A	N/A

geolocation_city	String	Cidade do cliente	N/A	N/A	Cidades do Brasil
geolocation_state	String	Estado do cliente	N/A	N/A	Estados do Brasil

4. Silver.dimitens_final:

Descrição: Detalha os itens dos pedidos realizados pelos clientes.

- freight_value: Valor do frete associado ao item (se um pedido tiver mais de um item, é preciso somar os valores do frete para verificar valor total)
- Uma ordem pode ter múltiplos itens, ou seja, esta tabela possui uma linha para cada item de cada pedido.
- Cada item pode ser atendido por um vendedor distinto.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
order_id	String	ID do pedido	N/A	N/A	N/A
order_item_id	Inteiro	Itens do pedido	N/A	N/A	N/A
product_id	String	ID do produto	N/A	N/A	Variações diversas de tipos de produtos
seller_id	String	Id do vendedor	N/A	N/A	N/A
shipping_limit_date	Timestamp	Data limite para envio do pedido	N/A	N/A	N/A
price	Decimal(10,2)	Preço do item	0	100000	N/A
freight_value	Decimal(10,2)	Preço do frete do item	0	1000	N/A

○

5. Silver.pagamentos_final:

Descrição: Contém informações sobre os pagamentos realizados para cada pedido.

- *payment_sequential: Sequência do pagamento no caso de múltiplos pagamentos para um pedido (um cliente pode pagar um pedido com mais de um método de pagamento. Se ele fizer isso, será criada uma sequência para acomodar todos os pagamentos.)

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
order_id	String	ID do pedido	N/A	N/A	N/A
payment_sequential	Inteiro	*	1	5	N/A
payment_type	String	Tipo de pagamento	N/A	N/A	'Cartão de crédito', 'boleto'

					bancário', etc
payment_installments	Inteiro	Número de parcelas	1	12	N/A
payment_value	Decimal(10,2)	Valor pago no pedido	0	100000	N/A

6. Silver.dimreviews_final:

Descrição: Contém as avaliações dos clientes após a entrega dos pedidos.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
review_id	String	ID da avaliação	N/A	N/A	N/A
order_id	String	ID do pedido	N/A	N/A	N/A
review_score	Inteiro	Nota da avaliação	1	10	N/A
review_comment_title	String	Título da avaliação	N/A	N/A	N/A
review_comment_message	String	Mensagem detalhada da avaliação	0	10000	N/A
review_creation_date	String (mantive como string, pois não iremos utilizar)	Data de criação da avaliação	N/A	N/A	N/A
review_answer_timestamp	String (mantive como string, pois não iremos utilizar)	Data e hora da resposta à avaliação	N/A	N/A	N/A

7. Silver.dimprodutos:

Descrição: Contém informações sobre os produtos vendidos na plataforma Olist.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
product_id	String	ID do produto	N/A	N/A	N/A
product_category_name	String	Categoria do produto	N/A	N/A	N/A
product_name_lenght	Inteiro	Número de caracteres do	2	50	N/A

		nome do produto			
product_description_lenght	Inteiro	Número de caracteres da descrição do produto	3	10000	N/A
product_photos_qty	Inteiro	Quantidade de fotos do produto	0	5	N/A
product_weight_g	Inteiro	Peso do produto em gramas	0	50000	N/A
product_length_cm	Inteiro	Comprimento em centímetros	0,1	500	N/A
product_height_cm	Inteiro	Altura em centímetros	0,1	500	N/A
product_width_cm	Inteiro	Largura em centímetros	0,1	500	N/A

8. Silver.dimvendedores:


Descrição: Contém informações sobre os vendedores da plataforma Olist.

Nome da coluna	Tipo de dado (dado bruto, sem transformações)	Descrição	Valor min. esperado	Valor max. esperado	Categorias possíveis
seller_id	String	ID do vendedor	N/A	N/A	N/A
seller_zip_code_prefix	Inteiro	Prefixo do CEP do vendedor	N/A	N/A	N/A
seller_city	String	Cidade do vendedor	N/A	N/A	Cidades brasileiras
seller_state	String	Estado do vendedor	N/A	N/A	Estados brasileiros

Catálogo de dados – Tabela camada Gold:

(*) Não foi alterado, permanece como na camada Silver.

Nome da coluna	Tipo de dado	Descrição	Valor mín. esperado	Valor máx. esperado	Categorias possíveis
order_id	String	*	*	*	*
Customer_id	String	*	*	*	*
Order_status	String	*	*	*	*
Total_items	Inteiro	Total de itens por pedido	1	100	N/A
Total_order_item_value	Decimal(10,2)	Valor total dos itens do pedido	0	100000	N/A
Total_freight_value	Decimal(10,2)	Valor total do frete do pedido	0	1000	N/A
Total_order_value	Decimal(10,2)	Valor total do pedido		100000	N/A
Order_purchase_timestamp_new	Date	*	*	*	*
order_approved_at_new	Date	*	*	*	*
order_delivered_carrier_date_new	Date	*	*	*	*
order_delivered_customer_date_new	Date	*	*	*	*
order_estimated_delivery_date_new	Date	*	*	*	*
payment_value	Decimal(10,2)	*	*	*	*
payment_type	String	*	*	*	*
product_id	String	*	*	*	*
price	Decimal(10,2)	*	*	*	*
product_category_name	String	*	*	*	*
seller_id	String	*	*	*	*
seller_city	String	*	*	*	*
customer_unique_id	String	*	*	*	*

 Silver.fatopedidos_final

 Silver.dimpagamentos

 Silver.dimitens

 Silver.DimProdutos

 Silver.DimVendedores

 Silver.DimClientes

Coleta e carga

Os arquivos CSV coletados no Kaggle (conforme exposto anteriormente) foram baixados manualmente e, após o download, os dados foram carregados para o DBFS em minha conta da plataforma Databricks Community Edition. Após a criação de um banco de dados com um comando SQL, o Spark foi utilizado para ler os arquivos CSV e registrá-los como tabelas no Databricks no database criado (“Bronze”). Nesta etapa, as tabelas foram criadas lendo

diretamente dos arquivos CSV, ou seja, a tabela continua sendo baseada em arquivo e não é otimizada para performance (carga bruta).

Para seguir na arquitetura em camadas, realizei a etapa de carga novamente, agora para a camada Silver. Nela, foi realizada a leitura dos CSVs com controle via Spark, incluindo schema, tratamento e transformação, para posteriormente salvar como uma tabela Delta, que possui vantagens, como o versionamento de dados, melhor performance e suporte a comandos, como “merge”, “update” e “delete”. Dessa forma, prosseguiu-se com a análise de possíveis tratamentos, como o de nulos, joins, conversão de tipos, filtros e renomeação.

Para finalizar, na camada “Gold” foi criada uma tabela a partir de Joins das tabelas da camada “Silver”, sendo elas a tabela fato de pedidos, DimPagamentos, DimItens, DimProdutos, DimVendedores e DimClientes. A criação desta tabela flat tem como objetivo otimizar a performance de leitura e análise dos dados, além de facilitar o seu consumo, permitindo assim responder de forma mais eficiente às perguntas definidas no início do projeto. Ainda assim, em caso de novas demandas analíticas, é possível recorrer à camada Silver, que mantém as tabelas organizadas em um modelo de esquema em floco de neve.

Análise

Qualidade de dados:

Após a carga dos dados na plataforma Databricks Community Edition e a criação das tabelas delta, foi realizada uma visualização inicial das tabelas de forma geral. Como o parâmetro “inferschema” foi configurado como “true” no Spark, as linhas do arquivo CSV foram analisadas automaticamente e seus tipos de dados supostamente apropriados foram retornados para cada coluna. No entanto, após análise, foi possível determinar que em etapas posteriores, na camada Silver, seriam necessárias algumas transformações do tipo de dado nas tabelas fato e a tabela denominada “DimReviews”.

Por consolidar a maioria das informações, a tabela de pedidos foi escolhida para ser a tabela fato e, armazenar dados quantitativos é característica deste tipo de tabela, contendo fatos e métricas de interesse. Por isso, foi decidido incluir à tabela quatro colunas de métricas simples, que são:

- Valor total dos itens do pedido: soma dos preços de cada item do pedido (cada linha pertence a 1 item, mesmo que seja do mesmo produto).
- Valor total do frete: como o valor do frete é calculado individualmente para cada item, é necessário somar o frete de cada item para obter o valor total.
- Valor total do pedido = produto total + Frete total.
- Total de itens por pedido.

Estas colunas foram acrescentadas afim verificar possíveis taxas extras e descontos.

Foi também realizada uma transformação na tabela de avaliações do cliente, na coluna “Review_score”, que constava como string, para o tipo de dado inteiro. Além disso, colunas que estavam como double nas tabelas de itens e pagamentos foram alteradas para o tipo decimal. Para finalizar esta etapa, as colunas relacionadas a data e hora do pedido na tabela fato foram formatadas para o tipo “Date”, mantendo apenas as informações de dia, mês e ano.

Um ponto importante a ser destacado é que, durante o processo, foi possível identificar que alguns casos de 'order_id' ausentes na tabela de itens possuem os status 'unavailable', 'canceled', 'unvoiced' ou 'created', ou seja, nenhum 'delivered'. Isso sugere que estes pedidos podem ter um comportamento diferente em relação à inclusão de itens na tabela DimItens. É possível que estes status específicos não tenham sido tratados corretamente ou não foram considerados relevantes para o preenchimento do conjunto de dados pertencentes a tabela DimItens.

Solução do problema:

Para esta etapa do projeto foi criada uma camada ouro ("*Gold*"), buscando responder as perguntas elencadas anteriormente. Como foco principal do estudo, a taxa de pedidos cancelados do conjunto de dados foi calculada, apresentando um valor de 0,63% que, apesar de pequena, equivale a 625 pedidos cancelados, que podem causar um impacto financeiro e logístico, além de uma experiência ruim ao cliente. Por isso, é preciso iniciar a investigação de porque os clientes ou empresas realizaram estes cancelamentos, se há alguma relação direta com o pagamento, falha na entrega ou insatisfação com o produto.

Continuando com as análises, foi possível identificar as categorias de produtos com os maiores números de cancelamentos entre 2016 e 2018. A categoria "utilidades_domesticas" se destacou, com 59 cancelamentos. Para responder a uma das questões levantadas no início deste trabalho, em seguida observou-se que, taxas de cancelamento altas são geralmente aquelas com um número menor de pedidos, com destaque para as categorias "pc_gamer", "portateis_cozinha_e_preparadores_de_alimentos" e "dvds_blu_ray". Já as categorias com grandes volumes de vendas e taxas de cancelamentos baixas, como "utilidades_domesticas" e "esporte_lazer", mostram um comportamento estável em relação aos cancelamentos.

Com o intuito de entender se estes cancelamentos estão correlacionados com as cidades dos vendedores, foi analisado quais cidades possuem o maior número de pedidos, o maior número de cancelamentos e a taxa de cancelamento. A cidade de São Paulo, como esperado, possui o maior número de pedidos, e 0,58% de taxa de cancelamento. A cidade de Pederneiras chamou atenção com 52,94% de taxa de cancelamento, dado que 9 de 17 pedidos foram cancelados. Para finalizar, Campinas possuiu uma taxa de cancelamento considerável, sendo ela 1,1%, com 16 cancelamentos e 1455 pedidos totais.

Os próximos questionamentos realizados estão relacionados a possíveis atrasos levarem ao cancelamento. Primeiramente foi verificado o tempo entre aprovação de pagamento e entrega ao cliente. Aqui já foi possível identificar que poucos clientes que receberam sua compra tiveram seu pedido cancelado posteriormente, visto que a grande maioria não possui uma data de entrega registrada. Isso indica que os cancelamentos, em sua maioria, não estão ligados à insatisfação com o produto, mas sim a outros fatores anteriores à entrega. As demais análises relacionadas a possíveis atrasos também demonstraram que não existe uma relação significativa entre atrasos na entrega e o cancelamento dos pedidos.

Para avaliar se os valores dos produtos influenciam na taxa de cancelamento, separei os produtos vendidos em quatro faixas de valores totais, sendo elas 0-50, 51-100, 101-200 e 201+1. Como resposta à pesquisa, a faixa de preços mais altos (201+) possui uma maior taxa de cancelamento, com 0,83%, o que poderia estar diretamente ligado a indecisões e

arrependimentos de compra devido ao preço mais elevado de produtos variados, já que foi notório que não há categorias específicas que devam preocupar.

Continuando com a questão financeira, foi verificado que o cartão de crédito é o tipo de pagamento que apresenta a maior quantidade de cancelamentos, o que é esperado dado o seu grande volume de pedidos (86.769). Sua taxa de cancelamento de 0,50% está entre as mais altas, juntamente com o voucher, que possui 0,46% (29 cancelamentos e 6.274 pedidos), e o boleto bancário, com 0,42%. O cartão de débito apresenta a menor taxa de cancelamento (0,35%), com uma quantidade baixa de cancelamentos (6) e pedidos (1.691). Isto pode significar que os consumidores que utilizam cartão de débito possam ter um maior controle sobre suas finanças, tornando-os menos propensos a cancelar pedidos, diferentemente do cartão de crédito, que pela sua flexibilidade pode levar as pessoas a agirem por impulsividade, resultando em arrependimento da compra, principalmente daquelas com valores mais elevados, como visto anteriormente.

Para finalizar este estudo, foi realizado um código que tem como principal objetivo identificar os clientes que realizaram cancelamentos em seus pedidos e calcular a taxa de cancelamento de cada um. Tivemos como retorno uma grande maioria de clientes com taxa de 100% de cancelamento, ou seja, um indicativo de que estes clientes não retornaram a comprar com nenhum vendedor. Isto pode estar diretamente ligado a falta de interesse ou impulsividade do consumidor ou até mesmo a um comportamento transacional sem compromisso, em que clientes realizam compras com o objetivo de testar o processo de compra ou se aproveitar de promoções. Para verificar este último mais afundo, verifiquei se os cancelamentos podem estar ligados a compras com descontos concedidos. Porém, a grande maioria dos pedidos cancelados não obteve descontos, podendo ser um incentivo para o cliente retornar a finalizar alguma outra compra, e até mesmo um motivador para futuros clientes, principalmente para aqueles que desejam comprar produtos com valores mais altos.

Conclusão sobre as análises e soluções dos problemas

Com base nas análises realizadas, é possível identificar fatores-chave que podem ser explorados e melhorados. O estudo apontou que os cancelamentos estão principalmente relacionados a comportamentos impulsivos dos clientes, especialmente em produtos de maior valor e quando o pagamento é feito por meio de métodos flexíveis, como o cartão de crédito. Além disso, a alta taxa de cancelamentos em Pederneiras merece uma investigação mais aprofundada, para identificar fatores locais que possam estar influenciando esses números e implementar melhorias específicas. Outro ponto relevante é que produtos com maior volume de vendas apresentam taxas de cancelamento mais baixas, o que sugere estabilidade. Para reduzir ainda mais a taxa geral de pedidos cancelados, campanhas de marketing ou ofertas especiais para esses produtos poderiam ser eficazes. Além disso, implementar programas de fidelidade e oferecer descontos para clientes com histórico de cancelamentos pode ajudar na fidelização, assim como incentivar o uso do débito como método de pagamento, oferecendo opções de financiamento sem juros.

Autoavaliação

O objetivo deste projeto foi analisar os fatores que influenciam o cancelamento de pedidos, buscando aplicar os conhecimentos adquiridos durante a Sprint de Engenharia de Dados da pós-graduação de Ciência de Dados e Analytics na PUC-Rio. Este projeto proporcionou uma excelente oportunidade para aprender ainda mais e desenvolver as habilidades acerca de banco de dados, Data Warehouse e linguagem SQL. Acredito que consegui responder às perguntas elencadas, conforme descrito nos requisitos do MVP, e fornecer possíveis soluções aos problemas encontrados. No entanto, o conjunto de dados escolhido pode ser muito mais explorado para melhor entendimento dos cancelamentos, como a análise das avaliações dos clientes. Além disso, um dashboard seria uma excelente adição para uma visualização mais clara dos resultados e identificação de padrões. Tenho a intenção de continuar otimizando este projeto e construindo novos para me desenvolver cada vez mais.