



PUC Minas

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação Lato Sensu em Analytics e Business Intelligence

RELATÓRIO TÉCNICO

DATASET BRAZILIAN E-COMMERCE BY OLIST

Luiz Felipe Pitta Gonçalves

São Paulo

2023

Sumário

1. INTRODUÇÃO	3
1.1. CONTEXTO.....	3
1.2. OBJETIVOS.....	3
1.3. PÚBLICO ALVO.....	3
2. MODELO DE DADOS	4
2.1. MODELO DIMENSIONAL	4
2.2. FATOS E DIMENSÕES.....	4
3. INTEGRAÇÃO, TRATAMENTO E CARGA DE DADOS	5
3.1. FONTES DE DADOS	5
3.2. PROCESSOS DE INTEGRAÇÃO E CARGA (ETL)	5
4. CAMADA DE APRESENTAÇÃO	7
4.1 DASHBOARD	7
4.1.1. PAINEL ESTRATÉGICO	7
4.1.2. PAINEL TÁTICO	8
4.1.3. PAINEL OPERACIONAL	9
4.2 ANÁLISES AVANÇADAS.....	11
5. REGISTROS DE HOMOLOGAÇÃO	12
6. CONCLUSÕES.....	13
7. LINKS	14

1. Introdução

1.1. Contexto

A empresa em questão utiliza diversas fontes de dados, incluindo bancos de dados transacionais, planilhas e sistemas legados. Essas informações estão dispersas e não estão integradas de forma centralizada. Além disso, os gestores e analistas enfrentam dificuldades em extrair insights relevantes dos dados devido à falta de uma ferramenta de análise adequada.

Nesse contexto, o projeto de desenvolvimento do painel de controle surge como uma solução para superar esses desafios. O objetivo é criar um ambiente visualmente intuitivo e de fácil acesso, no qual os usuários possam acompanhar os indicadores-chave de desempenho da empresa, analisar tendências, identificar padrões e realizar análises avançadas para embasar as decisões estratégicas e táticas.

1.2. Objetivos

O projeto está sendo desenvolvido com o objetivo de atender às necessidades da organização e satisfazer os seguintes objetivos:

- Melhorar o monitoramento e controle de indicadores-chave de desempenho.
- Fornecer uma visão estratégica, tática e operacional dos dados relevantes.
- Facilitar a tomada de decisões baseadas em dados.
- Minimizar falhas na integridade dos dados.
- Garantir um processo mais robusto e inteligente.

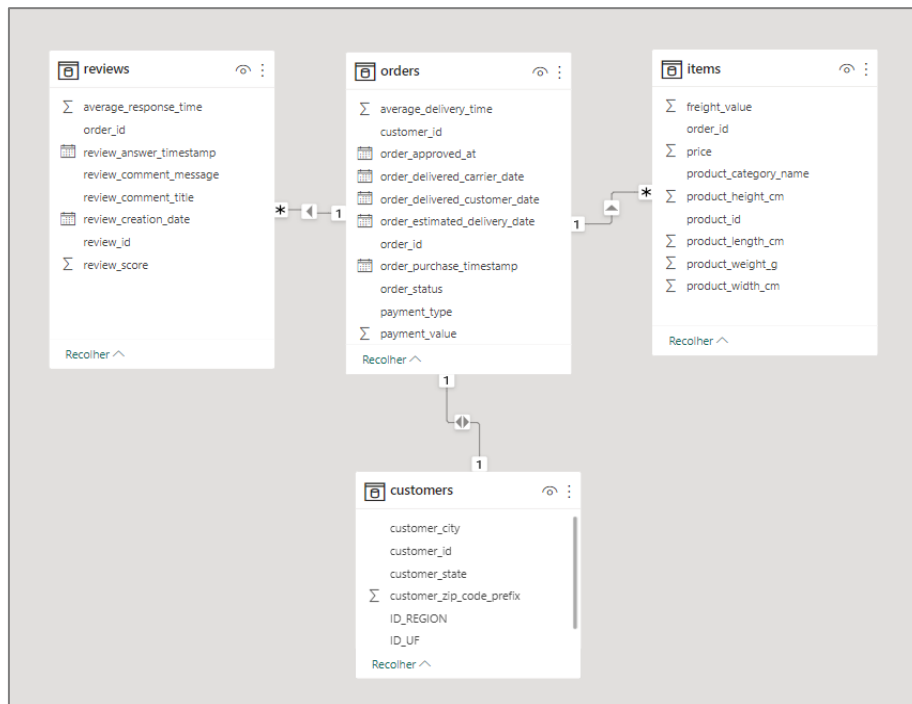
1.3. Público alvo

O público-alvo que utilizará a solução é composto por gestores, analistas e tomadores de decisão da organização.

2. Modelo de Dados

2.1. Modelo Dimensional

Para a construção do dashboard, foi utilizado a estrutura de relacionamentos entre tabelas a seguir:



2.2. Fatos e Dimensões

No projeto, foram identificadas as seguintes tabelas fatos e dimensões:

- Tabela Fato: A tabela Fato principal é denominada "Orders". Essa tabela contém informações sobre as transações de pedidos realizadas, como detalhes do pedido, datas, quantidades, valores e outras métricas relevantes. Ela serve como o ponto central para análises e relatórios relacionados aos pedidos efetuados.
- Tabelas Dimensão: Foram identificadas três tabelas de dimensão que fornecem contextos adicionais para enriquecer as análises:
 - "Reviews": Essa tabela de dimensão contém informações relacionadas às avaliações e feedback dos clientes sobre os pedidos. Inclui dados como a classificação do produto, comentários dos clientes e datas das avaliações.
 - "Items": A tabela de dimensão "Items" contém detalhes sobre os produtos incluídos nos pedidos, como informações de descrição, categorias, preços, entre outros. Essa dimensão permite analisar o desempenho dos itens e sua contribuição para as vendas e avaliações.

- "Customers": A tabela de dimensão "Customers" armazena informações sobre os clientes, como identificadores únicos, dados demográficos, localização. Essa dimensão permite segmentar e analisar o comportamento dos clientes, identificando padrões de compra, fidelidade e impacto nos pedidos.

3. Integração, Tratamento e Carga de Dados

3.1. Fontes de Dados

Todos os dados utilizados neste projeto foram extraídos de um dataset público disponível no Kaggle. Os dados podem ser encontrados no seguinte link: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>.

3.2. Processos de Integração e Carga (ETL)

No processo de ETL, foram implementadas melhorias para tornar o processo totalmente automático e otimizado. Essas melhorias foram estruturadas em diferentes etapas:

- Automação do download de fontes de dados: Para agilizar e simplificar o processo de obtenção dos dados, foi desenvolvida uma solução automatizada que realiza o download direto das fontes de dados do Kaggle. Isso elimina a necessidade de intervenção manual para obter os arquivos.
- Importação e análise exploratória: Após o download dos arquivos, foi implementada uma etapa de importação dos dados para o sistema. Além disso, uma análise exploratória mais robusta foi realizada para obter uma compreensão mais completa das estruturas e características dos dados. Isso permite identificar possíveis problemas ou padrões relevantes.
- Tratamento e refinamento dos dados: Para garantir a precisão e eficiência dos dados, uma série de operações de tratamento e refinamento foram aplicadas. Essas operações incluíram a remoção de duplicatas, preenchimento de valores faltantes, padronização de formatos e limpeza de outliers. O objetivo era obter dados mais enxutos e consistentes para as etapas subsequentes do projeto.

Com um processo automatizado e otimizado, o projeto pode se beneficiar de informações mais precisas e confiáveis para a tomada de decisões embasadas em dados. Para enriquecer ainda mais o entendimento, foram incluídas abaixo algumas imagens ilustrativas que representam visualmente as etapas mencionadas:

- Automação do download de fontes de dados

```
Download dataset

1 # Download dataset Kaggle
2 %pip install kaggle
3 !kaggle datasets download -d olistbr/brazilian-e-commerce

[ ] Python

1 # Caminho para captura do arquivo .zip
2 caminho_arquivo_zip = os.path.join(os.getcwd(), 'brazilian-e-commerce.zip')
3
4 # Dicionário para armazenar os dataframes
5 dataframes = {}
6
7 with zipfile.ZipFile(caminho_arquivo_zip, 'r') as zip_ref:
8     # Lista dos nomes dos arquivos CSV dentro do zip
9     nomes_arquivos_csv = [nome for nome in zip_ref.namelist() if nome.endswith('.csv')]
10
11 # Ler cada arquivo CSV em um dataframe separado
12 for nome_arquivo in nomes_arquivos_csv:
13     with zip_ref.open(nome_arquivo) as arquivo_csv:
14         # Obter o nome do arquivo sem a extensão .csv
15         nome_variavel = nome_arquivo[:-4]
16
17         # Ler o arquivo CSV em um dataframe
18         df = pd.read_csv(arquivo_csv)
19
20         # Adicionar o dataframe ao dicionário usando o nome do arquivo como chave
21         dataframes[nome_variavel] = df

[ ] Python
```

- Importação e análise exploratória

```
Análise exploratória geral

1 # Atribuir dataframes à variáveis
2 customers = dataframes['olist_customers_dataset']
3 geolocation = dataframes['olist_geolocation_dataset']
4 items = dataframes['olist_order_items_dataset']
5 payments = dataframes['olist_order_payments_dataset']
6 reviews = dataframes['olist_order_reviews_dataset']
7 orders = dataframes['olist_orders_dataset']
8 products = dataframes['olist_products_dataset']
9 sellers = dataframes['olist_sellers_dataset']
10 category = dataframes['product_category_name_translation']

[46] Python

1 # Visualizar dfs
2 for nome_variavel, df in dataframes.items():
3     print(f"Nome do dataframe: {nome_variavel}")
4     print(f"Dimensões: ", df.shape)
5     display(df.head(3))

[47] Python
```

- Tratamento e refinamento dos dados

```
Transformações nos Dataframes

> customers_dataset
1 cell hidden ...

> products_dataset
1 cell hidden ...

> items_dataset
1 cell hidden ...

> reviews_dataset
2 cells hidden ...

> payments_dataset
1 cell hidden ...

> orders_dataset
2 cells hidden ...
```

4. Camada de Apresentação

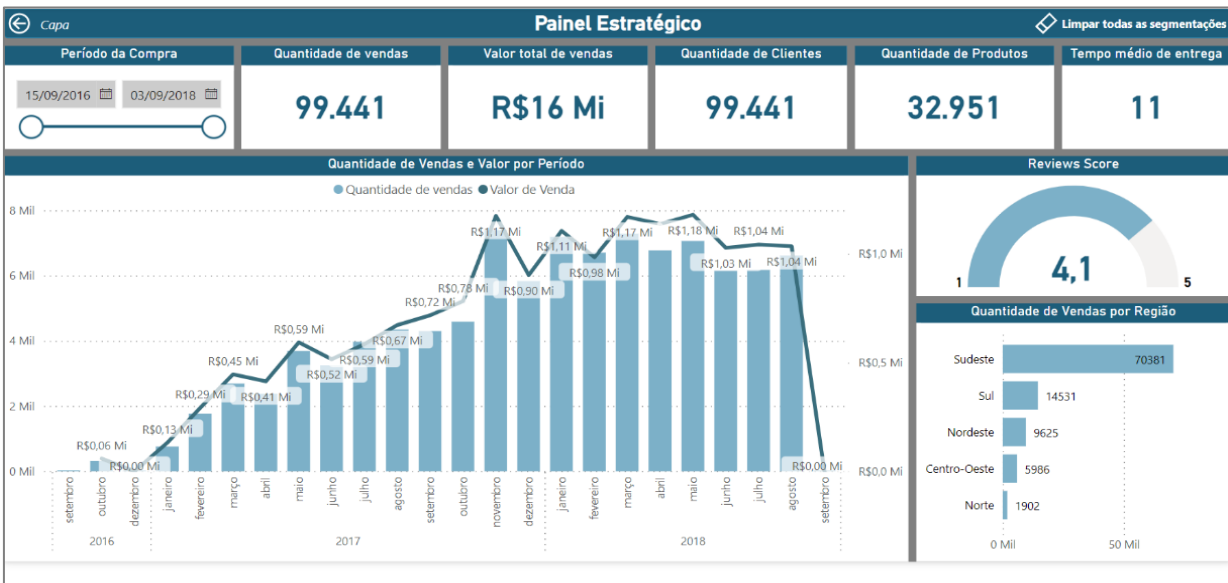
4.1 Dashboard

As relações entre as tabelas são estabelecidas por meio das seguintes chaves:

- Tabela "Orders" > "Reviews": Chave "order_id"
- Tabela "Orders" > "Items": Chave "order_id"
- Tabela "Orders" > "Customers": Chave "customer_id"

4.1.1. Painel Estratégico

- Dashboard:



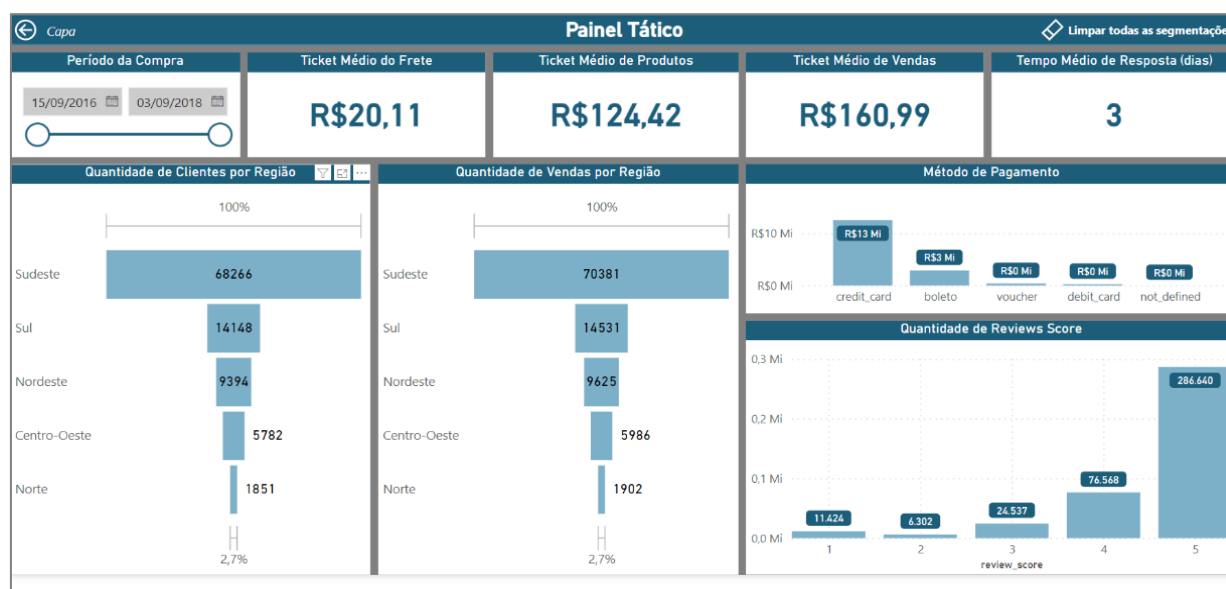
- Descrição dos indicadores:

Indicador	Dimensões	Cálculos	Tipo de visual	Descrição	Tabelas Fonte
Período de compra	Order_approved_at	-	Segmentação de dados	Este indicador permite segmentar os dados com base no período de compra.	Orders
Quantidade de vendas	Order_id	Contagem distinta	Cartão	Exibe a quantidade total de vendas realizadas	Orders
Valor total de vendas	Payment_value	Soma	Cartão	Exibe o valor total das vendas.	Orders
Quantidade de Clientes	Customer_id	Contagem distinta	Cartão	Exibe a quantidade total de clientes distintos.	Customers
Quantidade de Produtos	Product_id	Contagem distinta	Cartão	Exibe a quantidade total de produtos distintos vendidos.	Items
Tempo Médio de Entrega	Average_delivery_time	Média	Cartão	Exibe o tempo médio de entrega dos pedidos.	Orders

Quantidade de Vendas e Valor por Período	Order_approved_at, Order_id, payment_value	Contagem e Soma	Gráfico de colunas empilhadas e linha	Exibe a quantidade de vendas e o valor total das vendas por período.	Items, Orders
Reviews Score	Review_score	Mínimo, Média e Máximo	Indicador	Exibe as pontuações mínima, média e máxima das avaliações.	Reviews
Somatório de Vendas por Região	Region, order_id	Contagem	Gráfico de barras clusterizado	Exibe a soma de vendas por região geográfica.	Customers, Items

4.1.2. Painel Tático

- Dashboard:



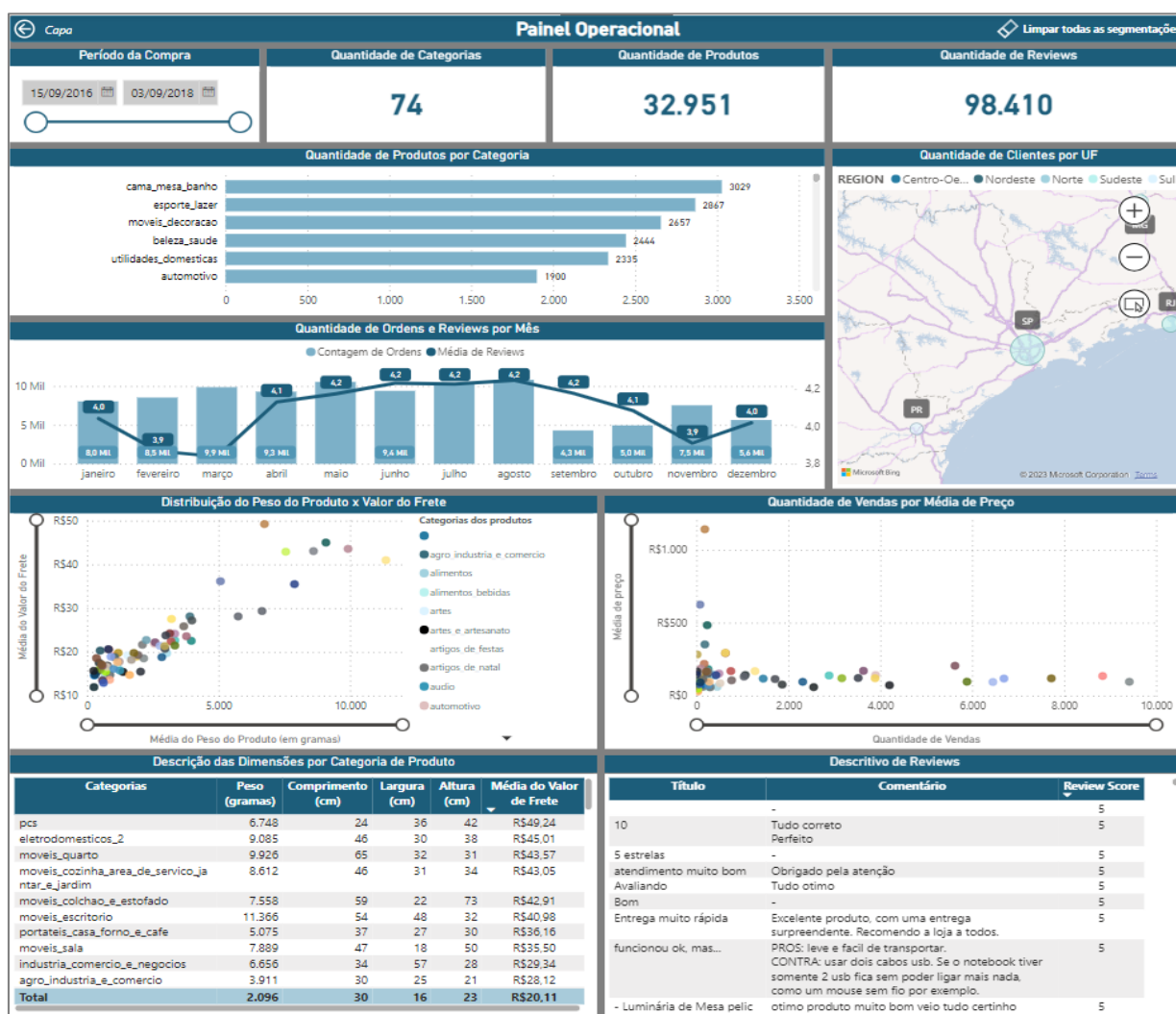
- Descrição dos indicadores:

Indicador	Dimensões	Cálculos	Tipo de visual	Descrição	Tabelas Fonte
Período de compra	Order_approved_at	-	Segmentação de dados	Este indicador permite segmentar os dados com base no período de compra.	Orders
Ticket Médio do Frete	Freight_value	Média	Cartão	Exibe o valor médio do frete por compra.	Items
Ticket Médio de Produtos	Price	Média	Cartão	Exibe o valor médio dos produtos por compra.	Items
Ticket Médio de Vendas	Payment_Value	Média	Cartão	Exibe o valor médio das vendas por compra.	Orders
Tempo Médio de Resposta(dias)	average_response_time	Média	Cartão	Exibe o tempo médio de resposta por compra.	Reviews

Quantidade de Clientes por Região	Region, Customer_id	Contagem	Funil	Exibe a quantidade de clientes por região.	Customers ,Orders
Quantidade de Vendas por Região	Region, Order_id	Contagem	Funil	Exibe a quantidade de vendas por região.	Customers ,Items
Método de Pagamento	Payment_type, payment_value	Soma	Gráfico de colunas clusterizado	Exibe a soma dos valores dos pagamentos por método de pagamento.	Orders
Quantidade de Reviews Score	Review_Score	Contagem	Gráfico de colunas empilhado	Exibe a quantidade de avaliações por pontuação.	Reviews

4.1.3. Painel Operacional

- Dashboard:

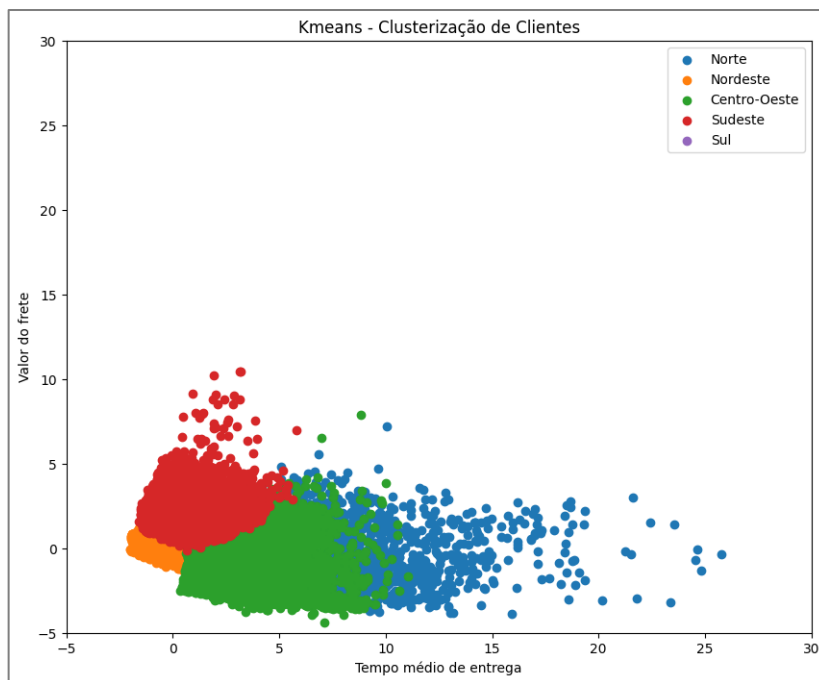


- Descrição dos indicadores:

Indicador	Dimensões	Cálculos	Tipo de visual	Descrição	Tabelas Fonte
Período de compra	Order_approved_at	-	Segmentação de dados	Este indicador permite segmentar os dados com base no período de compra.	Orders
Quantidade de Categorias	Product_category_name	Contagem	Cartão	Exibe a quantidade total de categorias de produtos.	Items
Quantidade de Reviews	Review_id	Contagem	Cartão	Exibe a quantidade total de reviews.	Reviews
Quantidade de Produtos por Categoria	Product_category_name, product_id	Contagem distinta	Gráfico de barras empilhado	Exibe a quantidade de produtos por categoria.	Items
Quantidade de Ordens e Reviews por Período	Order_approved_at, order_id, review_score	Contagem distinta e Média	Gráfico de colunas empilhado e linha	Exibe a quantidade de orders e o review score médio por período	Orders, Reviews
Quantidade de Clientes por UF	Customer_state, Region, customer_id	Contagem	Mapa	Exibe a quantidade de clientes por UF.	Orders, Customers
Distribuição do Peso do Produto x Valor do Frete	Product_weight_g, freight_value, product_category_name	Média	Gráfico de dispersão	Exibe a distribuição do peso dos produtos pelo valor médio do frete.	Items
Quantidade de Vendas por Média de Preço	Product_category_name, order_id, price	Contagem, Média	Gráfico de dispersão	Exibe a quantidade de vendas pelo valor médio de vendas.	Items
Descrição das Dimensões por Categoria de Produto	Product_category_name, freight_value, product_height_cm, product_length_cm, product_width_cm, product_weight_g	Média	Matriz	Exibe as dimensões dos produtos e o valor médio do frete.	Items
Descritivo de Reviews	Review_comment_title, review_comment_message, review_score	Primeiro e Média	Matriz	Exibe a descrição da avaliação e a nota do review.	Reviews

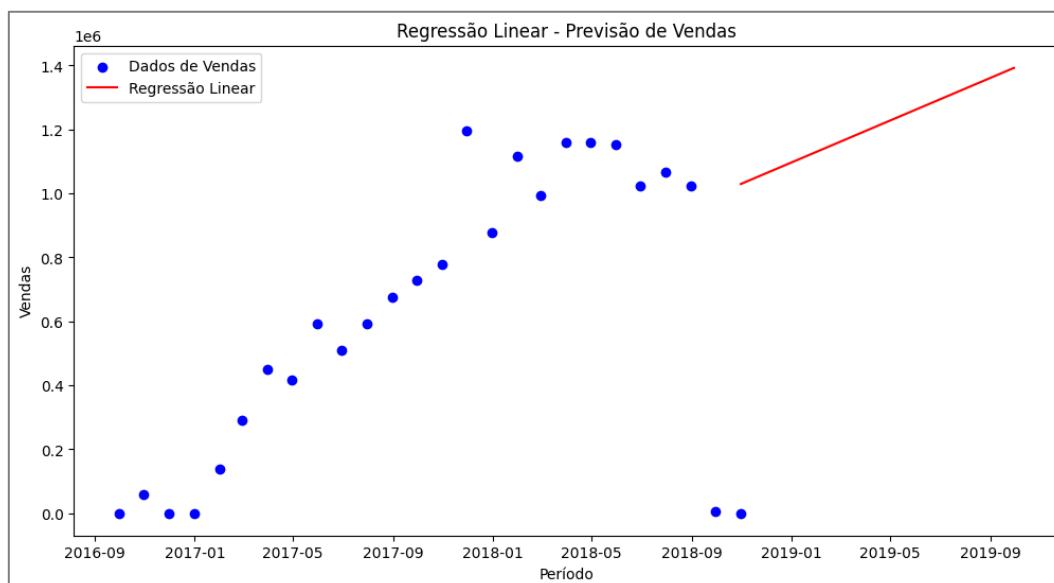
4.2 Análises avançadas

- Modelo de clusterização (K-means)



Conclusão: Foi observada uma correlação entre o custo do transporte e a duração média de entrega, indicando que as regiões Norte e Centro-Oeste têm um tempo de entrega mais longo. Além disso, ao aplicar o modelo K-means, não foram encontrados valores atípicos significativos no montante do frete. Conclui-se, portanto, que essas regiões enfrentam um maior tempo de entrega, enquanto os valores de frete se mantêm dentro da normalidade.

- Modelo de regressão linear



Conclusão: Ao projetarmos um período de 12 meses, notamos uma clara tendência de crescimento significativo, o que indica que a empresa está em um período de crescimento. Além disso, ao utilizar o modelo de regressão linear, confirmamos a associação direta entre o tempo decorrido e o

desempenho da empresa, reforçando a projeção de crescimento. Com base nessas análises, podemos concluir que a empresa está em um período promissor de expansão.

- Modelo de regras de associação

	Suporte	Categoria de produtos
1	0.095443	(cama_mesa_banho)
0	0.089555	(beleza_saude)
2	0.078244	(esporte_lazer)
3	0.067794	(informatica_acessorios)
4	0.065362	(moveis_decoracao)

Conclusão: Após a aplicação do modelo de regras de associação, cujo objetivo principal é identificar produtos frequentemente adquiridos em conjunto, ao analisar os resultados obtidos, constatamos que não há uma frequência significativa de compras conjuntas. Essa constatação nos leva a suspeitar que a base de dados possa ter sido manipulada.

5. Registros de Homologação

Homologações realizadas com o propósito de validar os dados.

```
Homologação

1 # Contagem distinta de ordens de venda
2 orders_trat.order_id.nunique()
[17] ✓ 0.0s Python
... 99441

1 # Soma do valor total de venda
2 orders_trat.payment_value.sum().round(2)
[18] ✓ 0.0s Python
... 16008872.12

1 # Contagem distinta de clientes
2 orders_trat.customer_id.nunique()
[19] ✓ 0.0s Python
... 99441

1 # Contagem distinta de produtos
2 items_trat.product_id.nunique()
[20] ✓ 0.0s Python
... 32951

1 # Ticket médio de vendas
2 orders_trat.payment_value.mean().round(2)
[21] ✓ 0.0s Python
... 160.99

1 # Ticket médio de frete
2 items_trat.freight_value.mean().round(2)
[22] ✓ 0.0s Python Python
... 20.11
```

6. Conclusões

- **Análise crítica**

Durante o processo de desenvolvimento e através das análises realizadas, foi possível identificar padrões e tendências relevantes para a organização. Isso incluiu informações sobre o comportamento dos clientes, desempenho dos produtos e eficiência dos processos de pedidos. Também foi possível notar que a organização enfrenta desafios na gestão de pedidos, incluindo atrasos na entrega, problemas de estoque e falta de visibilidade em tempo real sobre o status dos pedidos. Essas deficiências afetam a satisfação do cliente e a eficiência operacional. Algumas análises revelaram oportunidades de melhoria na experiência do cliente, como a personalização de ofertas, melhorias na comunicação e no suporte pós-venda. Essas ações podem levar a um aumento da fidelidade do cliente e a um maior número de recomendações.

- **Proposta de Intervenção**

Com base nos achados e conclusões obtidos neste estudo, recomenda-se as seguintes ações para aprimorar o desempenho da organização:

Aprimorar a gestão de pedidos: Recomenda-se a implementação de um sistema de gestão de pedidos integrado, capaz de oferecer monitoramento em tempo real do status dos pedidos. Essa solução permitirá a identificação ágil de problemas e a tomada de medidas corretivas imediatas. Ademais, é aconselhável investir em processos de logística eficientes a fim de mitigar a ocorrência de atrasos na entrega, o que contribuirá para uma experiência de compra satisfatória aos clientes.

Aperfeiçoar a experiência do cliente: Para elevar o nível de satisfação dos clientes, propõe-se a adoção de estratégias de personalização, as quais visam adequar ofertas e comunicações com base no perfil e histórico de cada cliente. Dessa forma, será possível fornecer um atendimento personalizado e relevante, ampliando a fidelidade e satisfação dos clientes. Além disso, é imprescindível investir em um suporte pós-venda eficiente, disponibilizando canais de atendimento ágeis e resolvendo prontamente as demandas e problemas dos clientes.

Estabelecer um ciclo de análise contínua: É fundamental estabelecer uma cultura de análise e monitoramento regular dos dados. Recomenda-se a realização de avaliações periódicas do desempenho da organização, permitindo a identificação de oportunidades de aprimoramento e antecipação de tendências. Ao embasar as decisões em dados confiáveis, a organização estará mais preparada para enfrentar os desafios do mercado e impulsionar o crescimento do negócio.

Essas ações propostas têm o potencial de impulsionar o desempenho da organização, promovendo uma gestão mais eficiente, aprimorando a experiência do cliente e consolidando uma cultura orientada por dados.

- **Lições aprendidas**

Durante o desenvolvimento deste projeto, foram adquiridas lições valiosas que podem ser aplicadas em futuros projetos. As principais lições aprendidas são as seguintes:

Preparação adequada dos dados: A preparação dos dados é um passo essencial para obter resultados confiáveis. Investir tempo e recursos na limpeza, refinamento e validação dos dados é fundamental para garantir a qualidade das análises. Essa etapa permite identificar e corrigir inconsistências, outliers e dados faltantes, proporcionando uma base sólida para as análises.

Análise contínua e adaptação: Os dados e as tendências estão em constante evolução. É fundamental adotar uma abordagem de análise contínua, acompanhando regularmente os resultados e monitorando as mudanças no ambiente. Através desse monitoramento, é possível identificar oportunidades de melhoria, antecipar tendências e tomar decisões mais informadas. A flexibilidade e capacidade de adaptação são essenciais para obter insights relevantes e manter a relevância das análises ao longo do tempo.

Integração de diferentes fontes de dados: A combinação de diferentes fontes de dados pode enriquecer as análises e fornecer insights mais completos. Ao integrar dados provenientes de diferentes sistemas e fontes, é possível obter uma visão mais abrangente e detalhada do negócio. Isso permite identificar correlações, padrões e tendências que podem passar despercebidos ao analisar apenas uma fonte de dados. A integração de dados é um desafio, mas os benefícios de uma visão holística compensam o esforço adicional.

Essas lições aprendidas são valiosas e podem ser aplicadas em outros projetos que envolvam análise de dados.

7. Links

Para facilitar o acesso a todos os arquivos utilizados no projeto, disponibilizo o link a seguir para o repositório no GitHub, onde estão armazenados os códigos-fonte, o dashboard, o arquivo README, o vídeo de apresentação e outros recursos relacionados: [GitHub - Brazilian Ecommerce](#)

Além disso, para visualizar diretamente o dashboard criado no Power BI, você pode acessar o seguinte link: [Dashboard PowerBI](#)