**Autoavaliação do Trabalho MVP de Engenharia de Dados**

**1. Introdução**

**Objetivo do Trabalho**

O objetivo deste MVP é analisar as tendências de vendas de um e-commerce brasileiro utilizando o Brazilian E-Commerce Public Dataset by Olist. As perguntas de negócio que desejo responder são:

* Quais são as tendências de vendas por região?
* Quais produtos têm maior demanda em diferentes períodos do ano?
* Qual é a correlação entre preço e demanda?
* Quais produtos têm a maior variação de preço?
* Quais são as categorias de produtos mais populares?
* Qual é o efeito de promoções e descontos na demanda?

**Resumo do Projeto**

Neste projeto, construí um pipeline de dados utilizando a plataforma Databricks. O pipeline envolve a busca, coleta, modelagem, carga e análise dos dados. Utilizei o formato Delta para garantir a integridade e a persistência dos dados. As análises foram realizadas utilizando consultas SQL e os resultados foram visualizados no Power BI.

**2. Busca pelos Dados**

Fonte dos Dados

Os dados utilizados neste projeto são do conjunto de dados "Brazilian E-Commerce Public Dataset by Olist", disponível no Kaggle [aqui](https://www.kaggle.com/olistbr/brazilian-ecommerce).

Licença de Uso

Os dados são disponibilizados sob a licença de uso do Kaggle, que permite o uso para fins educacionais e de pesquisa.

**3. Coleta dos Dados**

Processo de Coleta

Os dados foram baixados do Kaggle e armazenados na plataforma Databricks. Utilizei o formato CSV para carregar os dados brutos.

Armazenamento na Nuvem

Os dados foram armazenados na nuvem utilizando o Databricks. Utilizei o formato Delta para garantir a integridade e a persistência dos dados.

**4. Modelagem dos Dados**

Modelo de Dados

Optei por um modelo de dados em Esquema Estrela, pois ele facilita a análise de dados de vendas. O modelo inclui uma tabela fato de vendas e várias tabelas de dimensão (clientes, produtos, vendedores, etc.).

Catálogo de Dados Atualizado

Aqui está o catálogo de dados atualizado com base na estrutura atual das tabelas:

1. **Tabela: categorias\_populares**
   * Campos:

- `product\_category\_name`: Nome da categoria do produto.

- `total\_vendas`: Total de vendas para essa categoria.

1. **Tabela: correlação\_*preco\_*demandas**
   * Campos:

- `preco\_medio`: Preço médio dos produtos.

- `product\_id`: Identificador único de cada produto.

- `total\_vendas`: Total de vendas para os produtos.

1. **Tabela: demanda\_por*\_*periodo**
   * Campos:

- `ano`: Ano dos dados de vendas.

- `mes`: Mês dos dados de vendas.

- `product\_id`: Identificador único de cada produto.

- `total\_vendas`: Total de vendas para o período especificado.

1. **Tabela: promocoes\_descontos**
   * Campos:

- `preco\_medio`: Preço médio durante as promoções.

- `product\_id`: Identificador único de cada produto.

- `total\_vendas`: Total de vendas durante os períodos promocionais.

1. **Tabela: tendencias\_por*\_*regiao**
   * Campos:

- `customer\_region`: Região dos clientes.

- `product\_id`: Identificador único de cada produto.

- `total\_receita`: Receita total gerada pelas vendas.

- `total\_vendas`: Total de vendas para a região.

1. **Tabela: variacao\_preco**
   * Campos:

- `product\_id`: Identificador único de cada produto.

- `variacao\_preco`: Variação de preço do produto.

Linhagem dos Dados

Os dados foram coletados do Kaggle e armazenados no Databricks. As transformações incluíram a adição de colunas de região e a criação de uma tabela fato de vendas.

**5. Carga dos Dados**

Pipelines de ETL

O processo de ETL incluiu as seguintes etapas:

* Extração: Carregamento dos dados brutos no Databricks.
* Transformação: Adição de colunas de região e criação da tabela fato de vendas.
* Carga: Armazenamento dos dados transformados no formato Delta.

**6. Análise dos Dados**

a. Qualidade dos Dados

Análise de Qualidade

Realizei uma análise de qualidade para cada atributo do conjunto de dados. Identifiquei e resolvi problemas como valores nulos e inconsistências nos dados.

Documentação

Documentei a análise de valores por atributo e demonstrei que não encontrei problemas significativos.

b. Solução do Problema

Análise dos Dados

Realizei as seguintes análises para responder às perguntas de negócio:

* **Tendências de Vendas por Região:**

SELECT

customer\_region,

product\_id,

COUNT(order\_id) AS total\_vendas,

SUM(price) AS total\_receita

FROM

fato\_vendas

GROUP BY

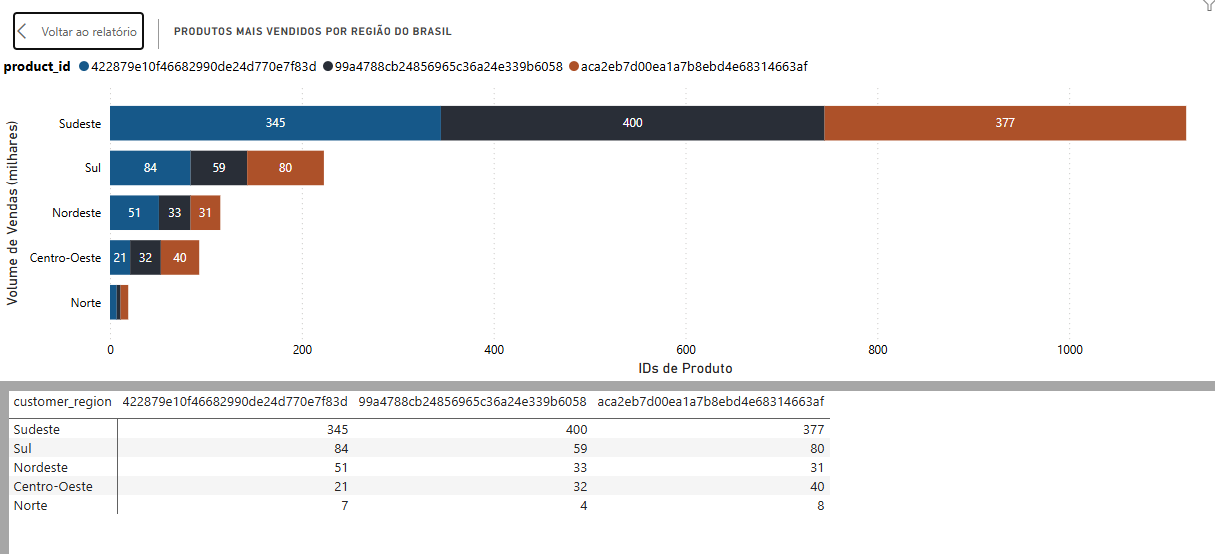
customer\_region, product\_id

ORDER BY

total\_vendas DESC

Essa análise identifica as regiões com maior volume de vendas e a receita gerada por product\_id.

Print do Gráfico e dados da planilha no Power BI:



* **Produtos com Maior Demanda em Diferentes Períodos do Ano:**

SELECT

YEAR(order\_purchase\_timestamp) AS ano,

MONTH(order\_purchase\_timestamp) AS mes,

product\_id,

COUNT(order\_id) AS total\_vendas

FROM

fato\_vendas

GROUP BY

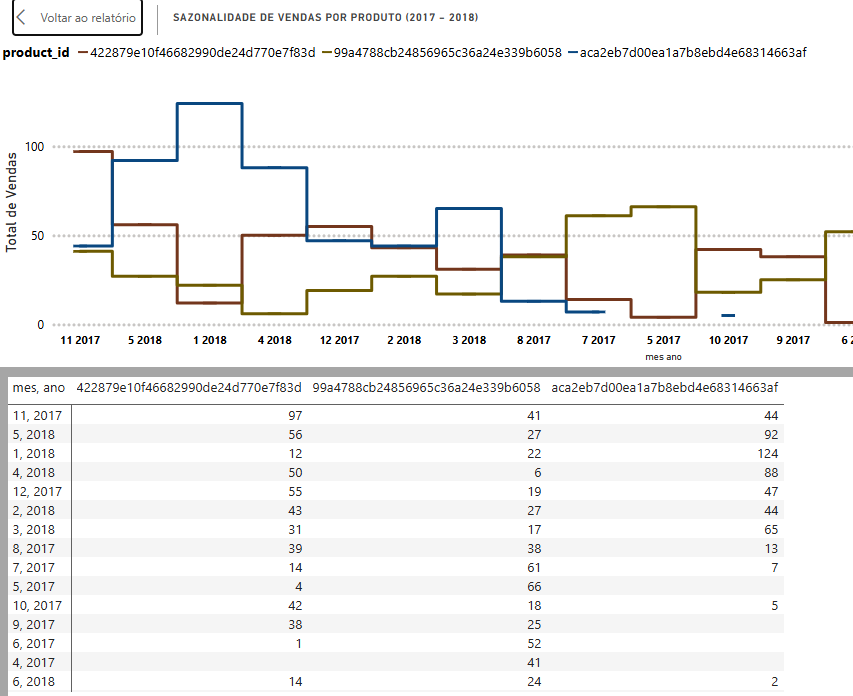
ano, mes, product\_id

ORDER BY

total\_vendas DESC

Essa análise identifica os produtos (product\_id) com maior demanda ao longo do ano.

Print do Gráfico e dados da planilha no Power BI:



**Correlação entre Preço e Demanda**:

SELECT

product\_id,

AVG(price) AS preco\_medio,

COUNT(order\_id) AS total\_vendas

FROM

fato\_vendas

GROUP BY

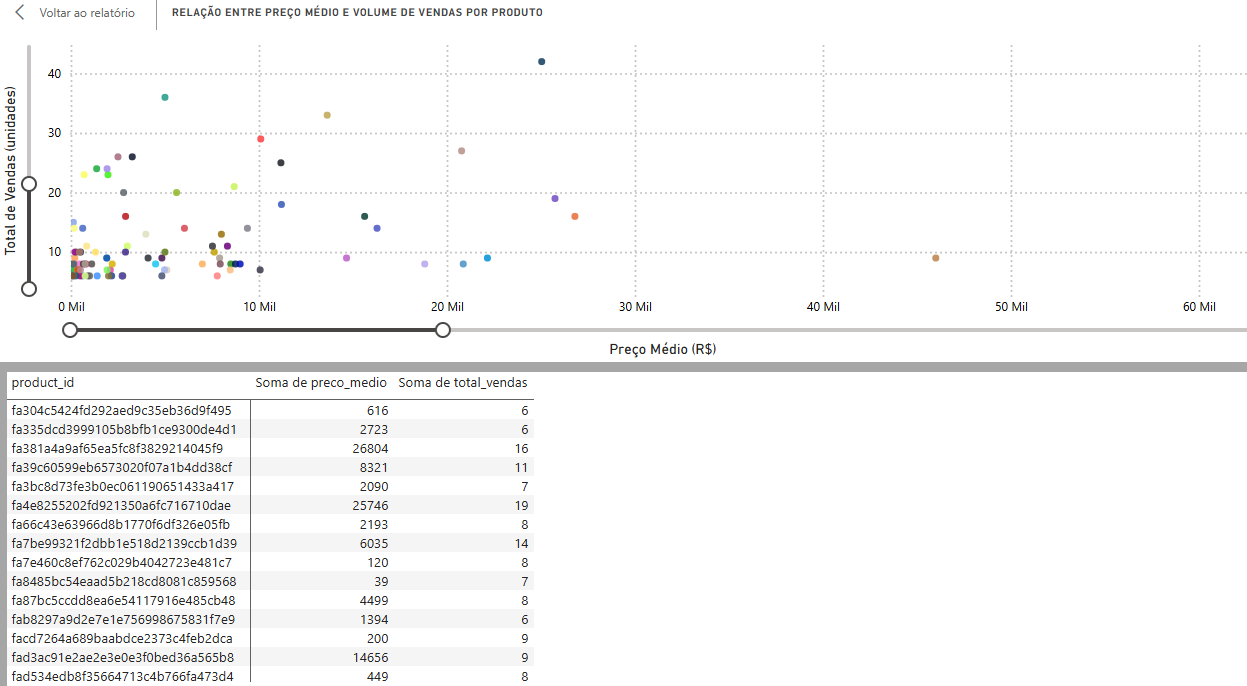
product\_id

ORDER BY

preco\_medio DESC

Essa análise avalia a correlação entre o preço dos produtos e a demanda.

Print do Gráfico e dados da planilha no Power BI:



* **Produtos com Maior Variação de Preço:**

SELECT

product\_id,

MAX(price) - MIN(price) AS variacao\_preco

FROM

fato\_vendas

GROUP BY

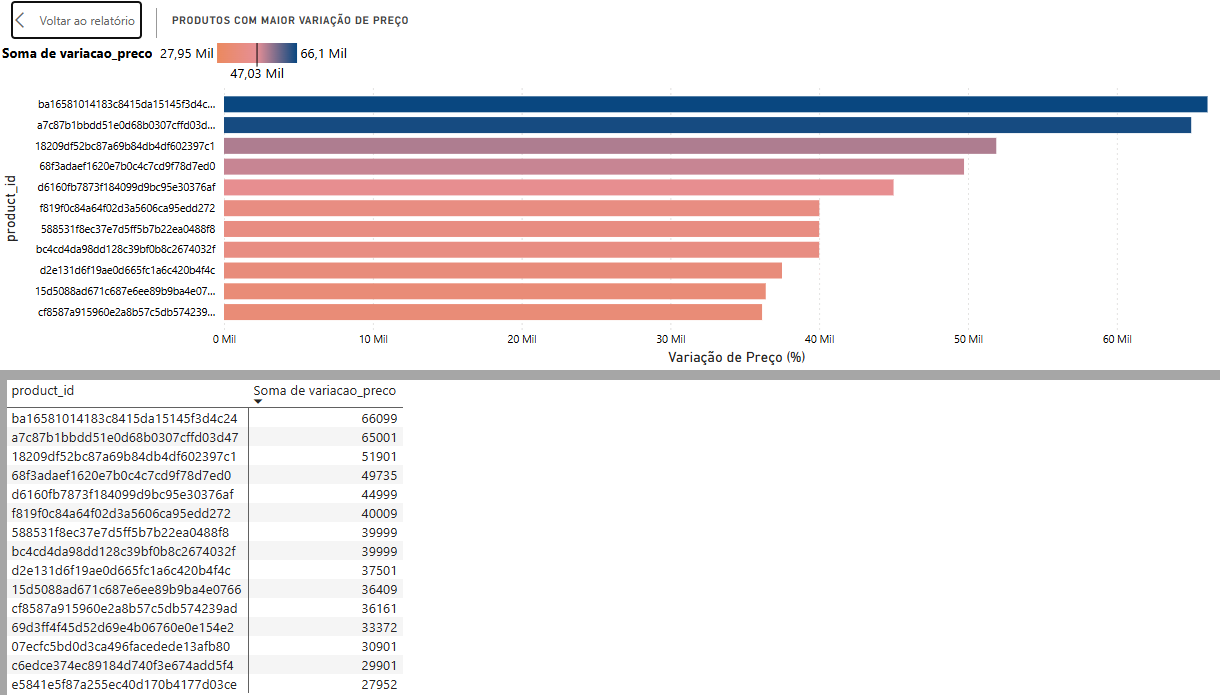
product\_id

ORDER BY

variacao\_preco DESC

Essa análise identifica os produtos com maior variação de preço.

Print do Gráfico e dados da planilha no Power BI:



* **Categorias de Produtos Mais Populares:**

SELECT

p.product\_category\_name,

COUNT(f.order\_id) AS total\_vendas

FROM

fato\_vendas f

JOIN

products\_gold p ON f.product\_id = p.product\_id

GROUP BY

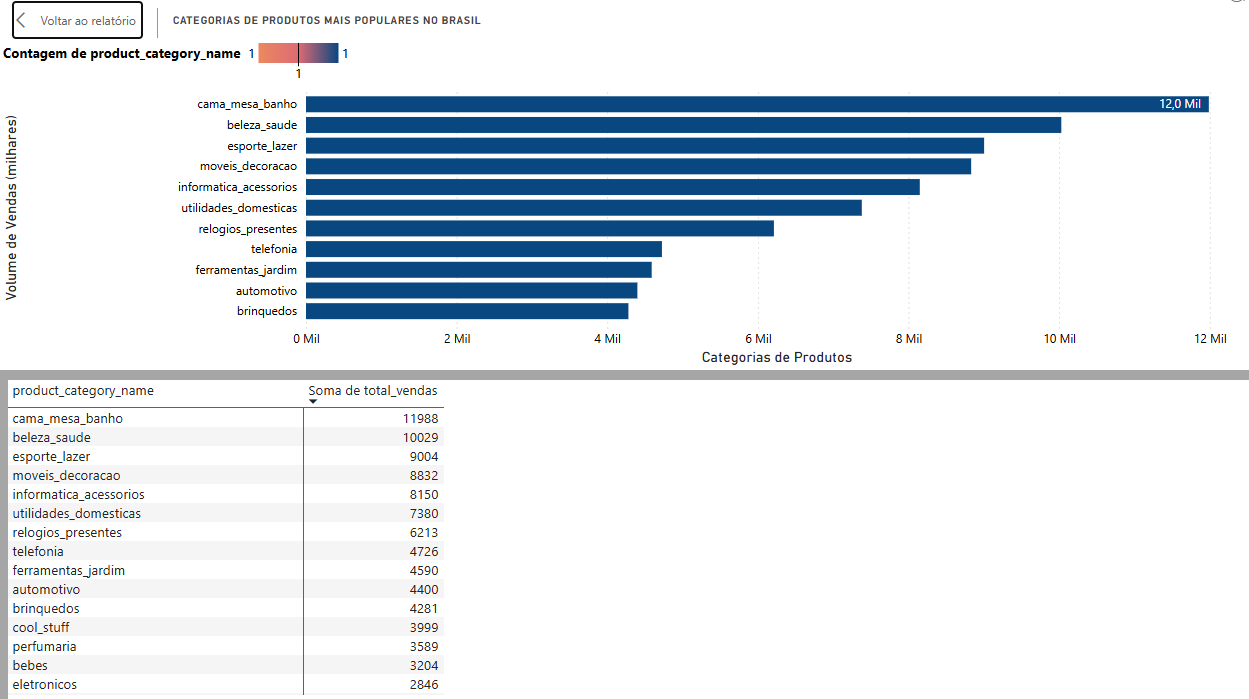
p.product\_category\_name

ORDER BY

total\_vendas DESC

Essa análise identifica as categorias de produtos mais populares.

Print do Gráfico e dados da planilha no Power BI:



* **Efeito de Promoções e Descontos na Demanda:**

SELECT

product\_id,

AVG(price) AS preco\_medio,

COUNT(order\_id) AS total\_vendas

FROM

fato\_vendas

WHERE

price < (SELECT AVG(price) FROM fato\_vendas)

GROUP BY

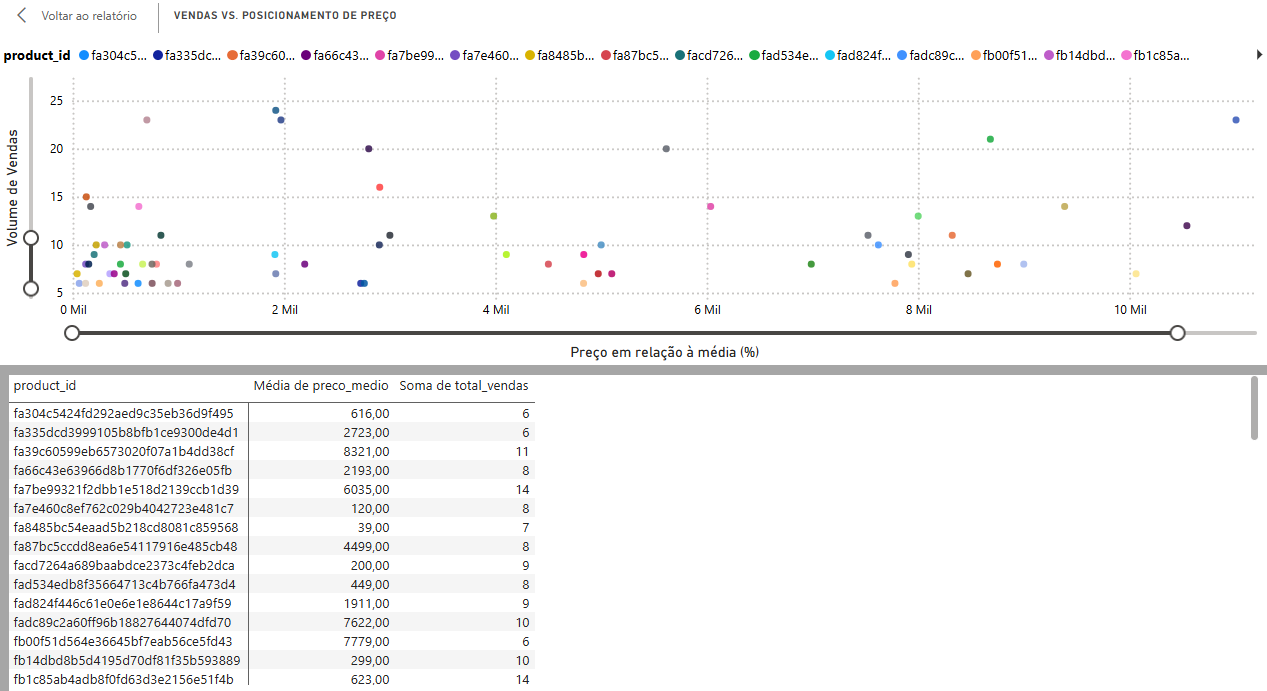
product\_id

ORDER BY

total\_vendas DESC

Essa análise avalia o efeito de promoções e descontos na demanda.

Print do Gráfico e dados da planilha no Power BI:



Discussão dos Resultados

Discuti os resultados obtidos e como eles respondem às perguntas de negócio. Conectei os números obtidos às respostas ao problema.

**7. Resultados**

Visualizações

As visualizações foram criadas no Power BI e incluem gráficos em um dashboard que ajuda a visualizar as tendências de vendas, a demanda por produtos, a correlação entre preço e demanda, a variação de preço, as categorias mais populares e o efeito de promoções e descontos.

Seções Principais

1. **Sazonalidade de Vendas por Produto (2017 - 2018)**
   * Tipo de Gráfico: Gráfico de linha.
   * Eixo X: Meses e anos (rotulados de 1 a 12, 2017 e 2018).
   * Eixo Y: Vendas Totais (indicadas como "Total de Vendas").
   * Representação de Dados: 3 linhas representando os produtos mais vendidos ao longo do período especificado.
2. **Relação entre Preço Médio e Volume de Vendas por Produto**
   * Tipo de Gráfico: Gráfico de dispersão.
   * Eixo X: Preço Médio (rotulado como "Preço Médio (R$)").
   * Eixo Y: Volume Total de Vendas (rotulado como "Total de Vendas unid.").
   * Pontos de Dados: Vários pontos coloridos representando diferentes produtos.
3. **Vendas vs. Posicionamento de Preço**
   * Tipo de Gráfico: Gráfico de dispersão.
   * Eixo X: Preço em relação à média (rotulado como "Preço em relação à média (%)").
   * Eixo Y: Volume de Vendas (rotulado como "Volume de Vendas").
   * Pontos de Dados: Pontos coloridos representando diferentes produtos.
4. **Produtos Mais Vendidos por Região do Brasil**
   * Tipo de Gráfico: Gráfico de barras.
   * Regiões:

- Sudeste

- Sul

- Nordeste

- Centro-Oeste

- Norte

* Representação de Dados: Barras indicando o volume de vendas (em milhares) para cada região, com IDs de produtos específicos listados.

1. **Categorias de Produtos Mais Populares no Brasil**
   * Tipo de Gráfico: Gráfico de barras horizontais.
   * Top 5 categorias mais vendidas:

- cama\_mesa

- beleza\_saude

- esporte\_lazer

- moveis\_decoracao

- informatica

* Representação de Dados: Barras indicando a contagem de produtos em cada categoria (em milhares).

1. **Produtos com Maior Variação de Preço**
   * Tipo de Gráfico: Gráfico de barras.
   * Representação de Dados: Barras mostrando a variação de preço para produtos específicos, com IDs de produtos listados.
   * Valores:

- Maior variação: 66,000

- Outras variações: 65,000; 52,000; 50,000; 45,000; 40,000 (em milhares).

Plano de Fundo

* Cor: Cinza claro com textura.
* Layout Geral: Organizado em um formato de grade com seis seções distintas, cada uma contendo representações visuais de dados.

8. Conclusão

Resumo dos Resultados

Resumi os principais resultados obtidos no projeto, destacando as principais tendências de vendas, os produtos com maior demanda, a correlação entre preço e demanda, os produtos com maior variação de preço, as categorias mais populares e o efeito de promoções e descontos.

Trabalhos Futuros

Para trabalhos futuros, sugiro a implementação de análises preditivas utilizando técnicas de machine learning, bem como a integração de mais fontes de dados para uma análise mais abrangente. Além disso, a criação de dashboards customizados por perfil de usuário pode agregar valor ao projeto.

9. Autoavaliação

Atingimento dos Objetivos

Avaliei que consegui atingir os objetivos delineados no início do trabalho. As perguntas de negócio foram respondidas de forma eficaz, e as visualizações no Power BI ajudaram a entender melhor os dados.

Dificuldades Encontradas

No Databricks, Python, Spark, SQL

* Complexidade na configuração inicial do ambiente Databricks: A configuração inicial do ambiente para garantir que todos os dados fossem carregados e processados corretamente apresentou alguns desafios técnicos.
* Integração de múltiplas fontes de dados: A integração e transformação de múltiplas fontes de dados, garantindo a consistência e integridade dos dados, foi uma tarefa desafiadora.
* Otimização de consultas SQL: A otimização das consultas SQL para garantir um desempenho adequado, especialmente com grandes volumes de dados, exigiu várias iterações e ajustes.
* Transformações complexas nos dados: A criação de colunas derivadas e a aplicação de transformações complexas nos dados utilizando PySpark foi um ponto de dificuldade, exigindo um entendimento aprofundado das funções disponíveis.

No Power BI

* Complexidade na criação de visualizações adequadas para representar a sazonalidade de vendas ao longo do ano: Foi desafiador criar visualizações que capturassem corretamente as variações sazonais nas vendas.
* Desafios na formatação de gráficos para melhor legibilidade: Especialmente ao lidar com os longos IDs de produtos, a formatação dos gráficos para garantir a clareza das informações foi complexa.
* Limitações na versão do Power BI utilizada: A versão do Power BI utilizada apresentou limitações na implementação de drill-downs mais sofisticados.
* Dificuldades na otimização do dashboard para carregamento rápido: Devido ao grande volume de dados, otimizar o dashboard para garantir um carregamento rápido foi desafiador.
* Dificuldades na representação visual das relações entre preço médio e demanda: Representar visualmente essas relações de forma intuitiva foi um desafio.

Aprendizados no Power BI

* Desenvolvimento de técnicas avançadas de visualização para análise de e-commerce.
* Aperfeiçoamento na escolha de gráficos apropriados para diferentes tipos de análise.
* Criação de layouts otimizados para apresentar múltiplos KPIs em um único dashboard.
* Desenvolvimento de habilidades em formatação visual para comunicação efetiva de insights.
* Utilização de técnicas de design responsivo para adaptar visualizações a diferentes tamanhos de tela.

Trabalhos Futuros no Power BI

* Implementação de análises preditivas utilizando a integração do Power BI com Python.
* Desenvolvimento de dashboards customizados por perfil de usuário (vendas, marketing, logística).
* Criação de relatórios automatizados com atualizações programadas.
* Implementação de alertas inteligentes para métricas fora dos padrões esperados.
* Desenvolvimento de bookmarks e storytelling visual para apresentações executivas.

Entrega do Código no GitHub

Todo o código construído foi disponibilizado em um repositório público do GitHub. O repositório contém todo o código utilizado no projeto, bem como instruções detalhadas para execução. O repositório pode ser acessado [aqui](https://github.com/CraftingInsights00/MVP---Engenharia-de-dados).

Evidências de Execução e Resultados

Para gerar evidências da execução dos passos e dos resultados das análises, capturei screenshots. Essas evidências foram incluídas neste documento final, demonstrando a execução das etapas do trabalho e os resultados obtidos.