Análise do Mercado de Medicamentos no Tratamento do Câncer de Próstata: Tendências e Previsões de venda

1. Objetivo

Este projeto tem como objetivo analisar as tendências de mercado de medicamentos utilizados no tratamento do câncer de próstata, com foco em casos metastáticos. A análise baseia-se em dados de vendas no Brasil dos últimos 24 meses, com o intuito de responder perguntas-chave de negócio e projetar a demanda futura desses medicamentos. Para isso, foi desenvolvido um pipeline de dados em nuvem, utilizando a plataforma Databricks Community, estruturado em camadas (Bronze, Silver e Gold) e modelado como um Data Warehouse no formato de esquema estrela.

Perguntas que se pretende responder:

1. Qual é a diferença entre o histórico e a previsão de vendas?
2. Quais são os medicamentos mais vendidos?
3. Quais estados ou cidades concentram o maior volume de vendas?
4. Quais canais (farmácia, hospitalar, outros) têm mais impacto nas vendas?
5. Qual é a previsão de vendas para os próximos 12 meses por produto e localidade?
6. Qual a tendencias de vendas por tipo de medicamento (marca, genérico, referência)?
7. Qual é a participação de medicamentos nacionais vs internacionais?

2. Coleta de Dados

Os dados utilizados foram extraídos da base da IQVIA, empresa na qual atuo, com devida autorização de uso. O arquivo, em formato CSV, foi carregado no ambiente Databricks por meio do DBFS (Databricks File System). O dataset contempla informações mensais de vendas de medicamentos voltados ao tratamento do câncer de próstata, segmentadas por estado, cidade, canal de distribuição, tipo e classe terapêutica dos medicamentos, entre outros atributos relevantes para análise.

3. Modelagem dos Dados

A modelagem dos dados foi realizada com base no arquivo CSV recebido. Esse arquivo foi processado e transformado em três tabelas distintas, a fim de viabilizar a construção de um esquema estrela, adequado para análises analíticas e visuais.

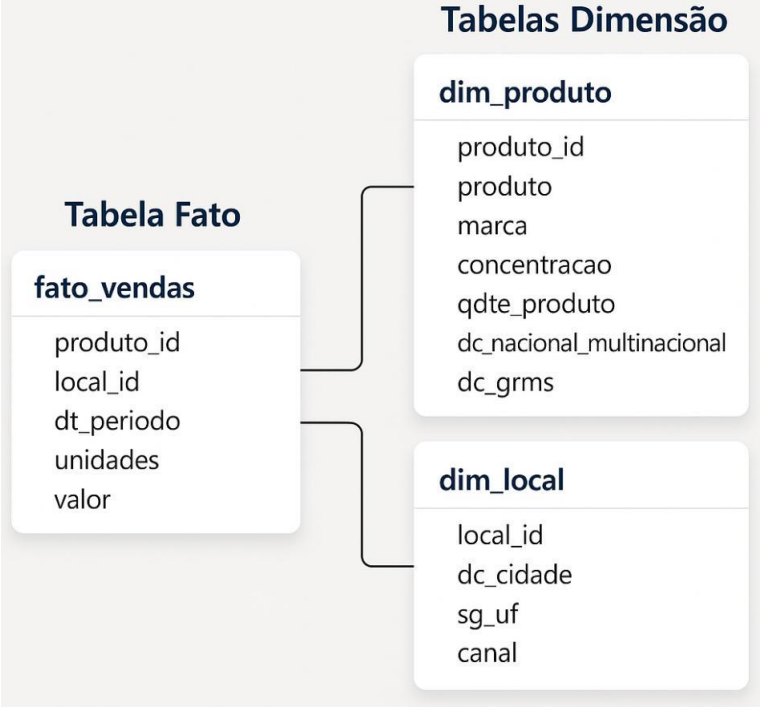
A estrutura do esquema é composta por:

Tabela Fato:

- fato_vendas: responsável por armazenar os dados transacionais de vendas mensais, contendo as colunas: produto_id, local_id, dt_periodo, unidades, valor.

Tabelas Dimensão:

- dim_produto: detalha os atributos dos medicamentos, com as colunas: produto_id, produto, marca, concentracao, qdte_produto, dc_nacional_multinacional, dc_grms.
- dim_local: descreve a localidade e o canal de venda, com os campos: local_id, dc_cidade, sg_uf, canal.



*Criado pela ferramenta Lucid

Catálogo de Dados

Este catálogo de dados descreve os campos presentes na tabela final da camada Gold, resultante da modelagem em esquema estrela implementada na camada Silver. A tabela Gold consolida os dados prontos para consumo analítico, incluindo informações de vendas históricas e de previsão, no qual foi utilizado o Modelo de Regressão em Gradient Boosted Trees (Árvores de Gradiente Aprimorado) para prever as unidades que serão vendidas para cada medicamento, marca, local, canal de venda e origem nos próximos 12 meses.

Cada coluna é documentada com sua respectiva função, tipo de dado, domínio esperado, categorias (quando aplicável) e observações relevantes, com o objetivo de garantir clareza.

Nome da Coluna	**Descrição**	**Tipo de Dado**	**Domínio / Intervalo Esperado**	**Categorias Possíveis**	**Fonte / Observações**
dt_periodo	Data das vendas realizadas e das previsões	Data	2023 - 2025	-	Derivado da data de venda
local_id	Identificador único do local	Inteiro	>= 0	-	Gerado na modelagem
produto_id	Identificador único do produto	Inteiro	>= 0	-	Gerado na modelagem
produto	Nome do medicamento	Texto	-	Nomes adaptados	Origem: dim_produto
marca	Marca do medicamento	Texto	-	Nomes adaptados	Origem: dim_produto
concentracao	Gramagem do medicamento	Texto	MG	-	Origem: dim_produto
qdte_produto	Unidade em cada caixa de medicamento	Inteiro	-	-	Origem: dim_produto
dc_nacional_multinacional	Tipo de laboratório (nacional ou multinacional)	Texto	-	Nacional, Multinacional	Origem: dim_produto
dc_grms	Tipo do medicamento	Texto	-	Genérico, Referência, Marca	Classificação commercial. Origem: dim_produto

dc_cidade	Cidade onde o medicamento foi ou tende a ser vendido	Texto	-	Nomes de cidades brasileiras	A partir da geolocaliza-ção na venda
sg_uf	Unidade federativa (UF) da venda e previsão	Texto	-	SP, RJ, MG, etc.	Sigla do estado
canal	Tipo do canal onde a venda ocorreu ou tende a acontecer	Texto	-	Farmácia, hospitalar, outros	Tipo de canal onde a venda foi registrada
origem	Se as vendas ocorreram ou é uma previsão	Texto	-	Histórico, Previsão	Criada na camada Gold
unidades	Quantidade de unidades vendidas	Inteiro	>= 0	-	Origem: base de vendas
valor	Valor das vendas	Decimal	>= 0	-	Valor monetário das vendas realizadas
unidades_previstas	Quantidade prevista (via modelo de regressão)	Decimal	>= 0	-	Resultado da previsão com GBTRegre-ssor

4. Carga (ETL)

O pipeline de ETL foi implementado no Databricks utilizando PySpark. As etapas incluíram:

Camada Bronze – Dados Brutos

Responsável pela ingestão dos dados originais, no formato CSV, com informações de vendas de medicamentos para **câncer de próstata metastático**, no período de **fev/2023 a jan/2025**. Os dados foram armazenados sem transformação para garantir a rastreabilidade.

Camada Silver – Tratamento e Estruturação

Nessa etapa, os dados foram limpos, padronizados e modelados no formato **estrela**, com a criação da tabela **fato_vendas** e dimensões como **produto e local**. Também foram geradas colunas como `produto_id`, `local_id`, preparando os dados para análise.

Camada Gold – Análise e Previsão

Com os dados prontos, aplicou-se o modelo preditivo de regressão **GBTRegressor (com PySpark)** para estimar a demanda futura dos medicamentos. As previsões foram armazenadas na camada Gold, baixada via CSV e imputada no **Power BI** para visualização e suporte das perguntas escolhidas.

5. Análise

a) Qualidade dos dados

Durante a transição dos dados da camada Bronze para a camada Silver, foram realizadas ações fundamentais para garantir a integridade, a consistência e a confiabilidade dos dados utilizados na análise. As principais etapas de limpeza e padronização aplicadas estão descritas a seguir:

1. Remoção de duplicatas:

Foi aplicada a função dropDuplicates() com o objetivo de eliminar registros duplicados, no qual durante o processo os dados já estavam limpos.

```
df_silver = df_bronze.dropDuplicates()
```

2.Conversão de tipos de dados:

Para assegurar a correta manipulação dos campos numéricos e temporais, foram realizados castings nas colunas:

- A coluna unidades foi convertida para o tipo integer.
- A coluna valor foi convertida para o tipo double.

```
df_silver = df_silver.withColumn("unidades", col("unidades").cast("integer"))
```

```
df_silver = df_silver.withColumn("valor", col("valor").cast("double"))
```

3.Padronização do campo de data (dt_periodo):

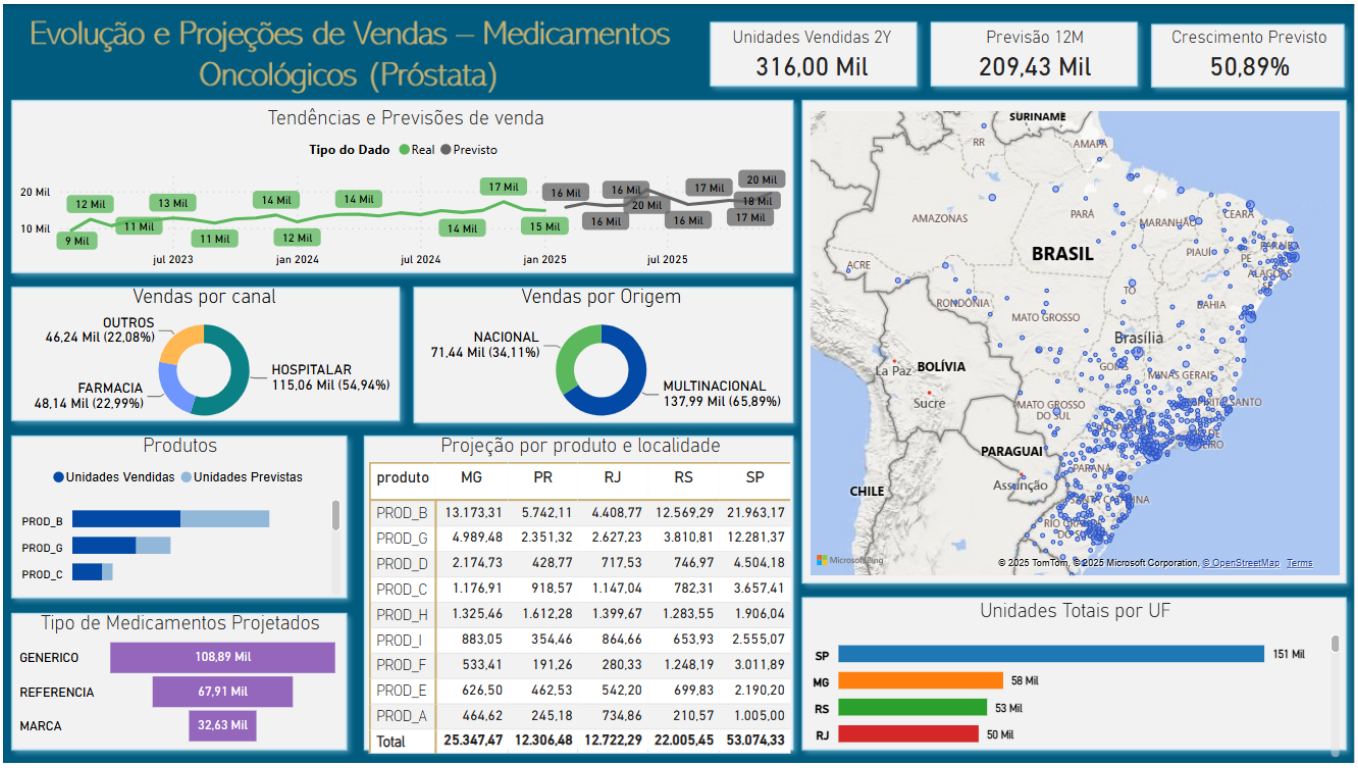
O campo dt_periodo, originalmente representado no formato "yyyymm" como string, foi transformado para o tipo date, com a data assumindo o primeiro dia do respectivo mês. Essa padronização permite maior flexibilidade para análises temporais, como agregações mensais ou comparações entre períodos.

```
df_silver = df_silver.withColumn(   "dt_periodo",  to_date(concat_ws("-", col("dt_periodo").substr(1, 4), col("dt_periodo").substr(5, 2), lit("01")), "yyyy-MM-dd"))
```

b) Solução do problema

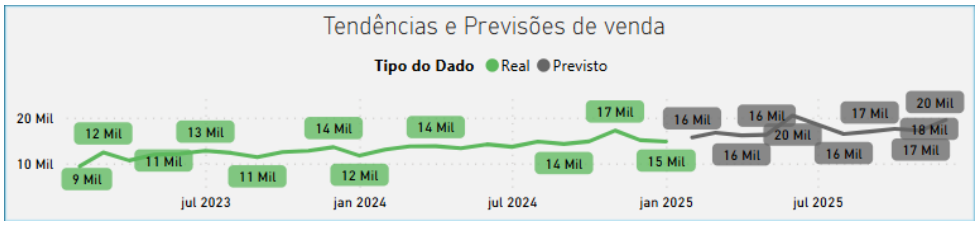
Foram gerados gráficos de análise no Power BI baseados na tabela final da camada gold, analisando a participação de mercado, tipos de medicamentos e a previsão de demanda.

Foi criada uma medida para corrigir uma anomalia detectada na previsão de vendas para julho/2025, que apresentava um pico atípico e incoerente com a tendência histórica. Para evitar distorções na análise, o valor desse mês foi ajustado com a média dos meses vizinhos (junho e agosto/2025). Nos demais meses, a medida realiza a soma normal entre unidades vendidas e previstas. Essa correção garante consistência visual, melhora a interpretação dos dados.



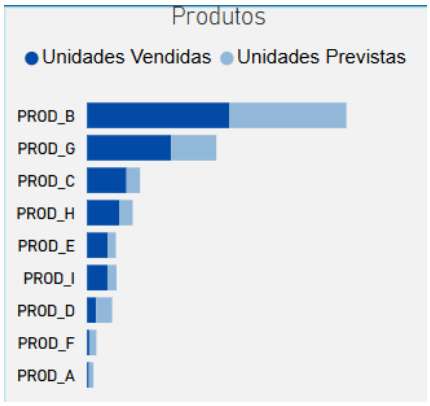
1. Qual é a diferença entre o histórico e a previsão de vendas?

Comparando os dados reais com os previstos no gráfico de linha, há crescimento previsto de 50,89%, sugerindo aceleração nas vendas neste ano.



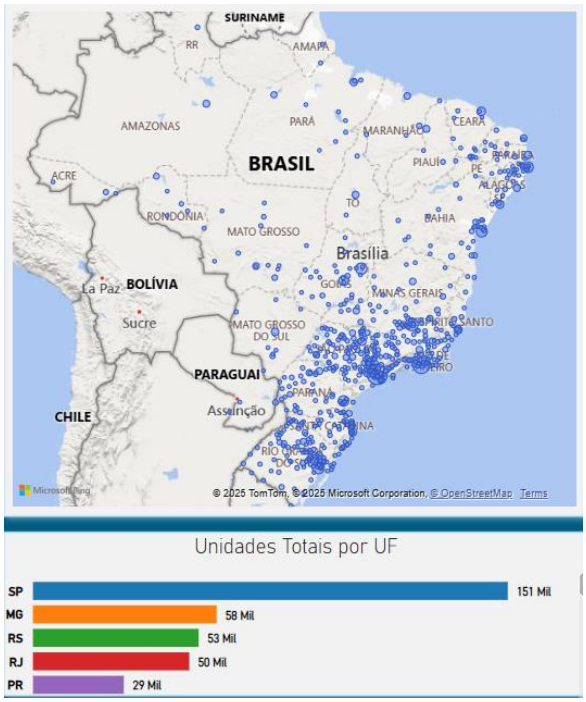
2. Quais são os medicamentos mais vendidos?

O medicamento PROD_B é o mais vendido, com grande destaque, seguido por PROD_G e PROD_C. Os demais possuem volumes significativamente menores.



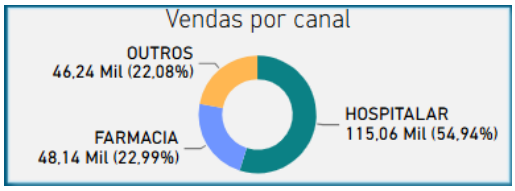
3. Quais estados ou cidades concentram o maior volume de vendas?

O estado de São Paulo (SP) lidera com 151 mil unidades, seguido por Minas Gerais (MG) com 58 mil, e Rio de Janeiro (RS) com 53 mil. O mapa geográfico também reforça a concentração no Sudeste e Sul do Brasil.



4. Quais canais (farmácia, hospitalar, outros) têm mais impacto nas vendas?

O canal Hospitalar é o mais impactante, com 54,94% das vendas, seguido por Farmácias e Outros.



5. Qual é a previsão de vendas para os próximos 12 meses por produto e localidade?

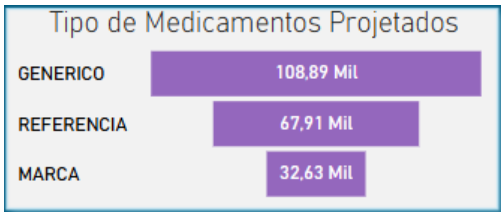
Há uma previsão total de 209,43 mil unidades, com destaque para SP (53 mil), MG (25 mil) e RS (22 mil). Os produtos com maior previsão são PROD_B e PROD_G, especialmente em SP.

Projeção por produto e localidade

produto	MG	PR	RJ	RS	SP
PROD_B	13.173,31	5.742,11	4.408,77	12.569,29	21.963,17
PROD_G	4.989,48	2.351,32	2.627,23	3.810,81	12.281,37
PROD_D	2.174,73	428,77	717,53	746,97	4.504,18
PROD_C	1.176,91	918,57	1.147,04	782,31	3.657,41
PROD_H	1.325,46	1.612,28	1.399,67	1.283,55	1.906,04
PROD_I	883,05	354,46	864,66	653,93	2.555,07
PROD_F	533,41	191,26	280,33	1.248,19	3.011,89
PROD_E	626,50	462,53	542,20	699,83	2.190,20
PROD_A	464,62	245,18	734,86	210,57	1.005,00
Total	**25.347,47**	**12.306,48**	**12.722,29**	**22.005,45**	**53.074,33**

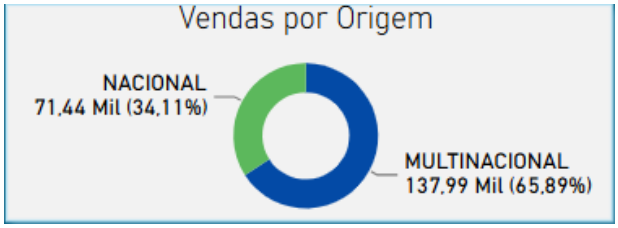
6. Qual a tendencias de vendas por tipo de medicamento (marca, genérico, referência)?

Medicamentos genéricos lideram com mais de 50% da previsão total, indicando preferência de mercado e potencial maior de distribuição.



7. Qual é a participação de medicamentos nacionais vs internacionais?

Os medicamentos de origem internacional lideram o mercado, sendo quase o dobro da participação nacional.



6. Autoavaliação

O projeto cumpriu todos os objetivos propostos, permitindo a análise de tendências e previsões de vendas de medicamentos para o tratamento do câncer de próstata. Entre os principais desafios enfrentados, destaca-se a implementação de modelos preditivos no ambiente Databricks Community, que exigiu testes com diferentes algoritmos até a escolha do GBTRegressor, por apresentar melhor desempenho.

Outro ponto de atenção foi a exportação da camada Gold em formato CSV para uso no Power BI, devido ao volume expressivo de dados. Foi necessário planejar cuidadosamente essa transição para garantir a integridade e a performance na visualização.

Além disso, identifiquei uma anomalia nas previsões de julho/2025, com um pico fora do padrão. Para manter a coerência com a tendência histórica, foi aplicado um ajuste utilizando a média dos meses adjacentes (junho e agosto/2025).

A organização em camadas (bronze, silver e gold) e a modelagem dimensional em esquema estrela foram bem-sucedidas, garantindo uma estrutura de dados limpa, eficiente e analiticamente robusta.

7. Evidências

Em anexo junto com essa documentação está:

- O notebook utilizado no Databricks com as etapas da pipeline;
- Screenshots do dashboard no Power BI e o próprio arquivo utilizado;
- Base CSV dos dados que utilizei, fornecidos pela IQVIA.