GRUPO 19

Richard Santos - RA 12923117888

Andrew Da Luz Borges - RA 1292212243

Erika Lacerda Vieira - RA 1292216836

Sumário

1. Definição do Problema

Este projeto tem como objetivo analisar dados históricos de vendas disponibilizados pelo Walmart na competição M5 Forecasting. Buscamos identificar padrões e tendências com foco em criar modelos preditivos de vendas diárias por loja e produto, para subsidiar estratégias comerciais, logísticas e de precificação.

2. Base de Dados Utilizada

A base de dados é composta pelos seguintes arquivos:
- sales_train_validation.csv: Vendas diárias de produtos por loja.
- calendar.csv: Contém informações temporais como feriados e eventos.
- sell_prices.csv: Histórico de preços por produto e loja.

As principais colunas incluem identificadores de produtos e lojas, datas no formato 'd_1' a 'd_1913', indicadores econômicos e datas especiais.

3. Importação e Limpeza de Dados

O código abaixo foi utilizado para importar os dados, transformar o dataset de vendas para o formato longo e tratar os dados ausentes:

```
[40] import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
```

Nova seção

```
[28] # ============================ #
     # 1. CARREGAMENTO DOS DADOS   #
     # ============================ #

     print("\n📥 Carregando datasets...")

     sales = pd.read_csv("/content/sales_train_validation.csv",
                         dtype={
                             'item_id': 'category', 'dept_id': 'category',
                             'cat_id': 'category', 'store_id': 'category',
                             'state_id': 'category'
                         })

     calendar = pd.read_csv("calendar.csv", dtype={'d': 'category'})
     sell_prices = pd.read_csv("/content/sell_prices.csv",
                               dtype={
                                   'store_id': 'category',
                                   'item_id': 'category',
                                   'wm_yr_wk': 'int16',
                                   'sell_price': 'float32'
                               })
```

```
📥 Carregando datasets...
```

[38]
```python
# ============================ #
# 2. LIMPEZA E TRANSFORMAÇÃO  #
# ============================ #

print("\n🧹 Limpando dados...")

# Remover duplicatas em sell_prices
dups = sell_prices.duplicated().sum()
if dups > 0:
    print(f"⚠️ Removendo {dups} duplicatas em sell_prices.")
    sell_prices = sell_prices.drop_duplicates()

# Transformar sales para formato longo
sales_long = pd.melt(sales_sample,
                     id_vars=['id', 'item_id', 'dept_id', 'cat_id', 'store_id', 'state_id'],
                     var_name='d',
                     value_name='sales')

# Otimizações
sales_long['d'] = sales_long['d'].astype('category')

# Merge com calendar
sales_long = sales_long.merge(calendar_reduced, on='d', how='left', validate='many_to_one')

# Preencher valores ausentes
event_cols = ['event_name_1', 'event_type_1', 'event_name_2', 'event_type_2']
sales_long[event_cols] = sales_long[event_cols].fillna("None")
sales_long['sales'] = sales_long['sales'].fillna(0)
sales_long['sales'] = sales_long['sales'].replace([float('inf'), -float('inf')], 0)
sales_long['sales'] = sales_long['sales'].astype('int32')

# Conversões finais
sales_long['date'] = pd.to_datetime(sales_long['date'], format='%Y-%m-%d')
for col in event_cols:
    sales_long[col] = sales_long[col].astype('category')
```

```
🧹 Limpando dados...
```

```python
# ============================ #
# 3. RELATÓRIO FINAL           #
# ============================ #

print("\n📊 Resumo de valores ausentes:")
print(sales_long.isnull().sum())

print("\n📦 Informações finais do DataFrame:")
print(sales_long.info(memory_usage='deep'))
```

```
📊 Resumo de valores ausentes:
id              0
item_id         0
dept_id         0
cat_id          0
store_id        0
state_id        0
d               0
sales           0
date            0
event_name_1    0
event_type_1    0
event_name_2    0
event_type_2    0
dtype: int64

📦 Informações finais do DataFrame:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19130000 entries, 0 to 19129999
Data columns (total 13 columns):
 #   Column        Dtype
---  ------        -----
 0   id            object
 1   item_id       category
 2   dept_id       category
 3   cat_id        category
 4   store_id      category
 5   state_id      category
 6   d             object
 7   sales         int32
 8   date          datetime64[ns]
 9   event_name_1  category
 10  event_type_1  category
 11  event_name_2  category
 12  event_type_2  category
dtypes: category(9), datetime64[ns](1), int32(1), object(2)
memory usage: 3.0 GB
None
```

4. ANALISE EXPLORATÓRIA DE DADOS E MODELAGEM

Gráfico: Vendas Totais por Dia

- O gráfico temporal das vendas totais diárias ao longo de cerca de cinco anos revela uma tendência geral de crescimento gradual no volume de vendas. Nota-se também uma forte sazonalidade, com flutuações regulares, bem como quedas pontuais marcantes — possivelmente relacionadas a feriados ou falhas na coleta de dados. O aumento consistente ao longo do tempo pode refletir tanto a expansão da operação do Walmart quanto uma adaptação bem-sucedida às demandas do mercado. Esses padrões temporais são cruciais para o desenvolvimento de modelos preditivos eficazes.

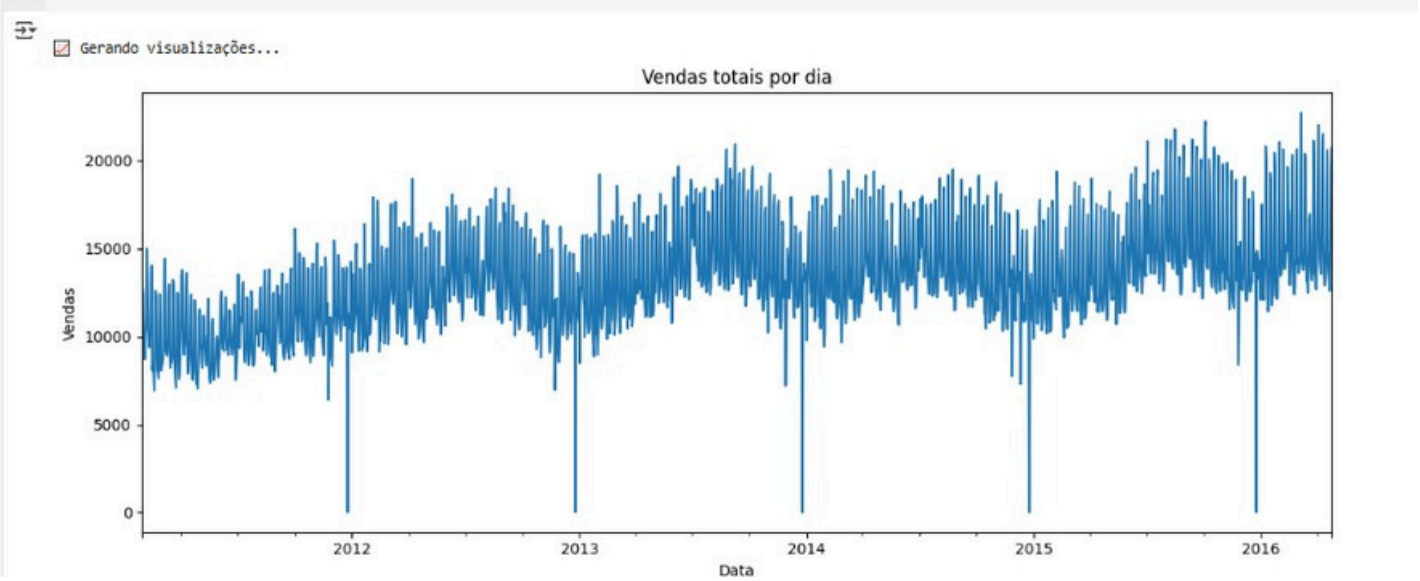


Gráfico: Top 10 Eventos (event_name_1)

- Este gráfico mostra os 10 principais eventos identificados no dataset, com destaque absoluto para a categoria “None”, representando datas sem eventos registrados. Observa-se que a grande maioria das datas não está associada a eventos específicos, o que sugere que as vendas no Walmart ocorrem predominantemente de forma regular, sem depender diretamente de eventos sazonais ou comemorativos. Contudo, também se destacam eventos como o início da Quaresma ("LentStart"), o "SuperBowl" e o "ValentinesDay", indicando potenciais impactos sazonais específicos sobre determinadas categorias de produtos.

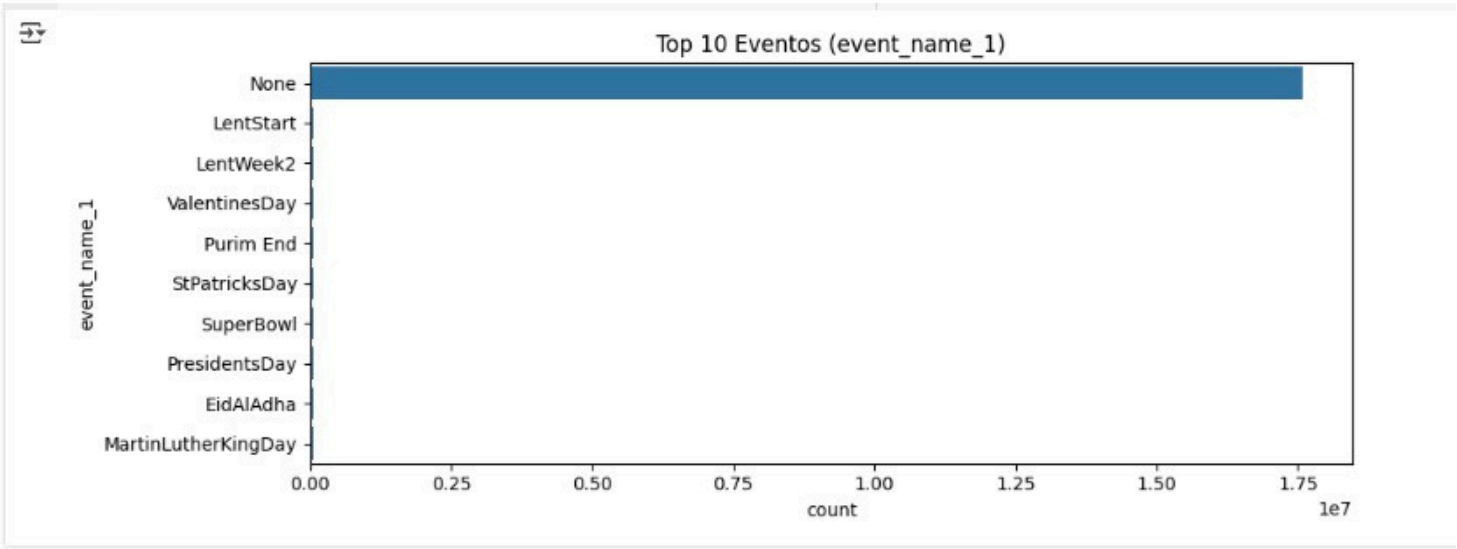


Gráfico: Vendas Totais por Estado

- O gráfico revela que a Califórnia (CA) lidera expressivamente o volume total de vendas, seguida pelo Texas (TX) e Wisconsin (WI). Este resultado sugere que as lojas da Califórnia desempenham um papel central no faturamento da rede Walmart, possivelmente devido à maior densidade populacional e ao número de unidades no estado. A diferença no volume de vendas entre os estados reforça a necessidade de estratégias comerciais diferenciadas, ajustadas conforme o perfil e a demanda regional.

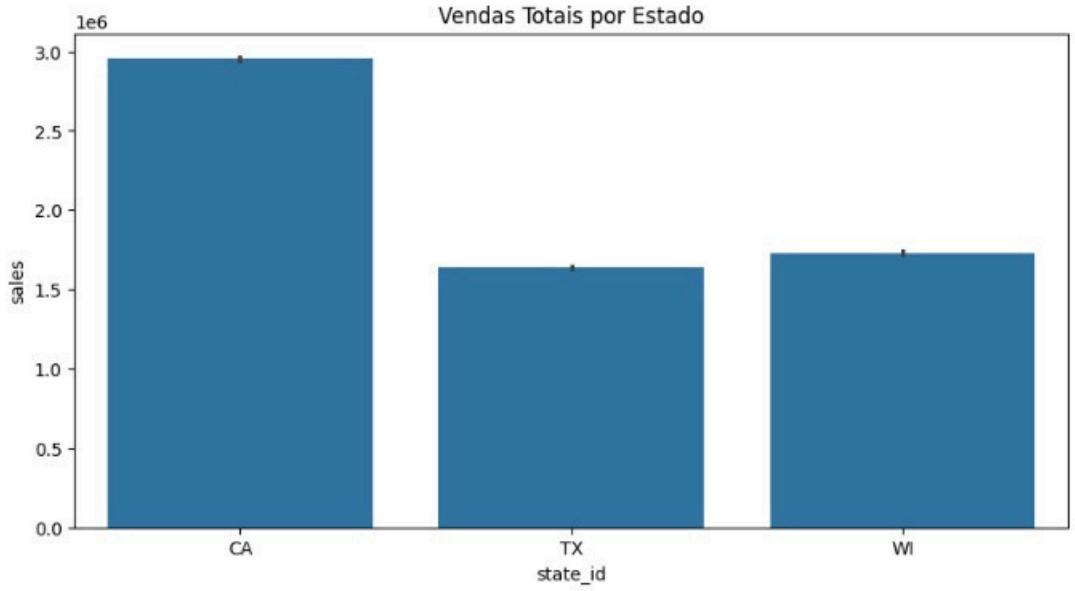
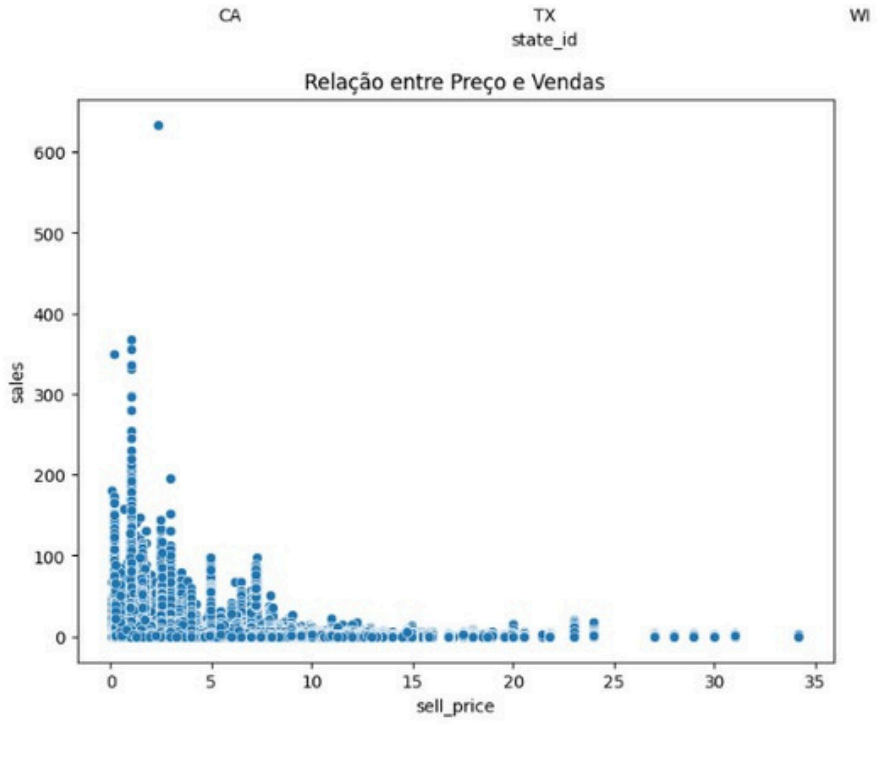


Gráfico: Relação entre Preço e Vendas

- Este gráfico de dispersão ilustra a relação inversa entre preço de venda (sell_price) e quantidade vendida (sales). Nota-se que a maioria das vendas ocorre com preços abaixo de US$ 10, enquanto produtos com preços superiores a US$ 20 são vendidos em menor quantidade. Esta distribuição confirma a lei da demanda, onde preços mais baixos tendem a estimular volumes de vendas maiores, enquanto produtos de maior valor são adquiridos com menor frequência, provavelmente por serem considerados itens de nicho ou com menor rotatividade.



Distribuição das Vendas

- A distribuição das vendas revela um padrão fortemente assimétrico à direita, com a maioria das observações concentradas em baixíssimos volumes de vendas, próximas de zero. Apenas uma fração muito pequena dos registros apresenta volumes de vendas elevados, com valores que chegam a ultrapassar 600 unidades. Este comportamento é típico de mercados de varejo com grande diversidade de produtos, onde a maioria apresenta baixa rotatividade, enquanto poucos itens são altamente demandados. Esta distribuição sugere a importância de estratégias diferenciadas: políticas de long tail para produtos com pouca saída, e campanhas promocionais ou reforço logístico para itens de alta demanda.

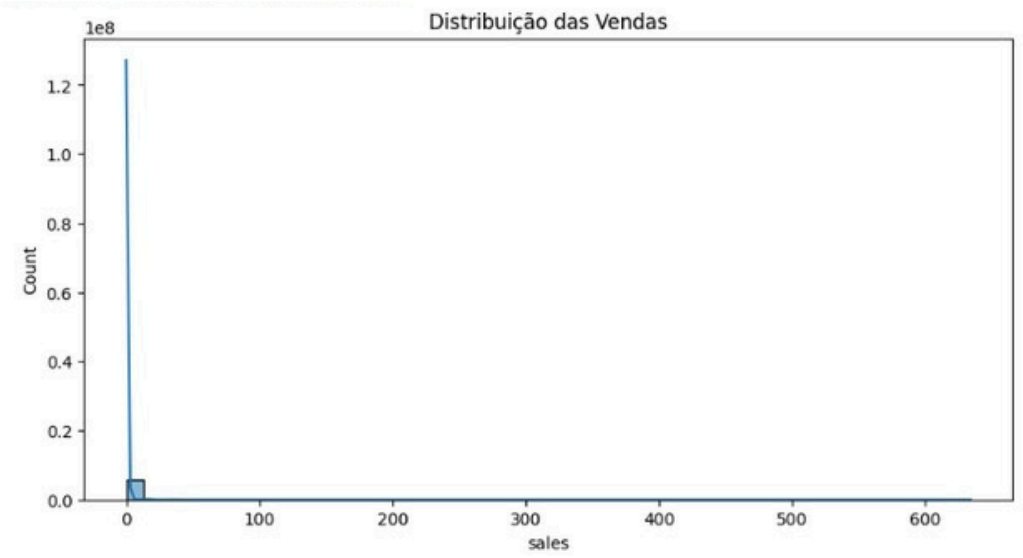


Gráfico: Boxplot de Vendas por Categoria

Este boxplot apresenta a distribuição das vendas dentro de cada uma das três principais categorias: “HOBBIES”, “HOUSEHOLD” e “FOODS”. Nota-se que a categoria de alimentos (FOODS) possui a maior dispersão, com diversos outliers acima de 600 unidades vendidas, demonstrando uma maior variabilidade e volume potencial de vendas. Já as categorias “HOBBIES” e “HOUSEHOLD” apresentam distribuições mais concentradas, com volumes de vendas mais modestos e menos dispersos. Esta análise evidencia que o setor alimentício não só lidera em vendas totais, mas também em volatilidade, o que exige uma gestão mais cuidadosa de estoques e reposições.

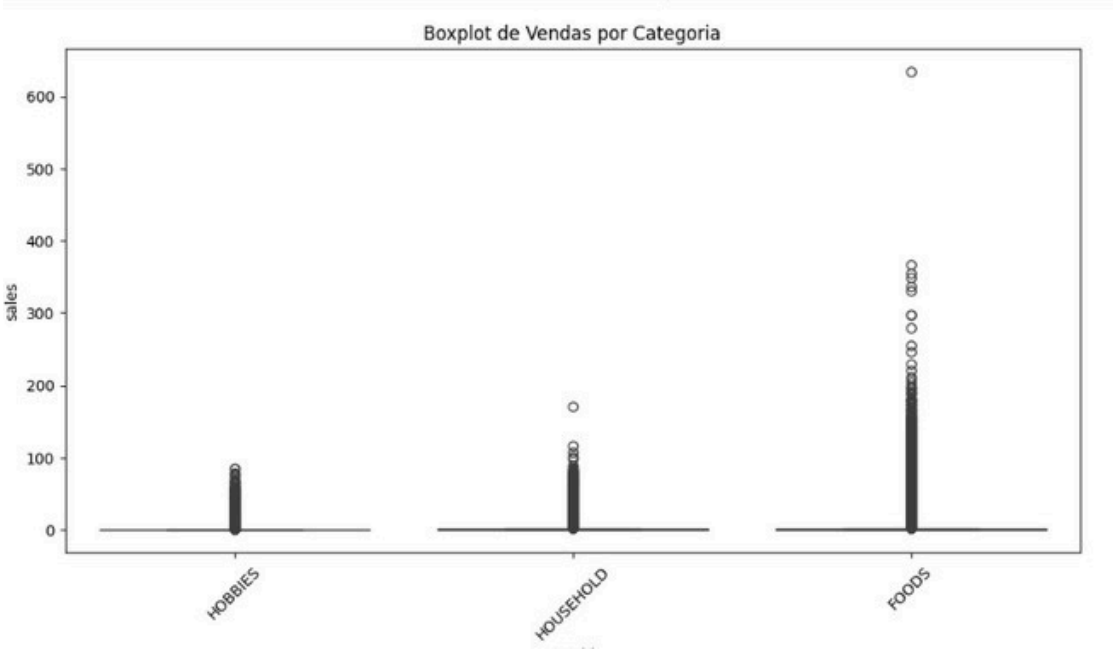


Gráfico: Vendas Médias por Dia da Semana

O gráfico destaca variações significativas no volume médio de vendas conforme o dia da semana. As vendas são relativamente constantes de segunda a quinta-feira, apresentando leve elevação às sextas. O destaque vai para os sábados e domingos, que registram os maiores volumes médios de vendas. Este comportamento sugere que o consumo é intensificado durante os fins de semana, possivelmente em função de maior disponibilidade de tempo dos consumidores ou promoções específicas, reforçando a importância de estratégias comerciais direcionadas a esses dias.

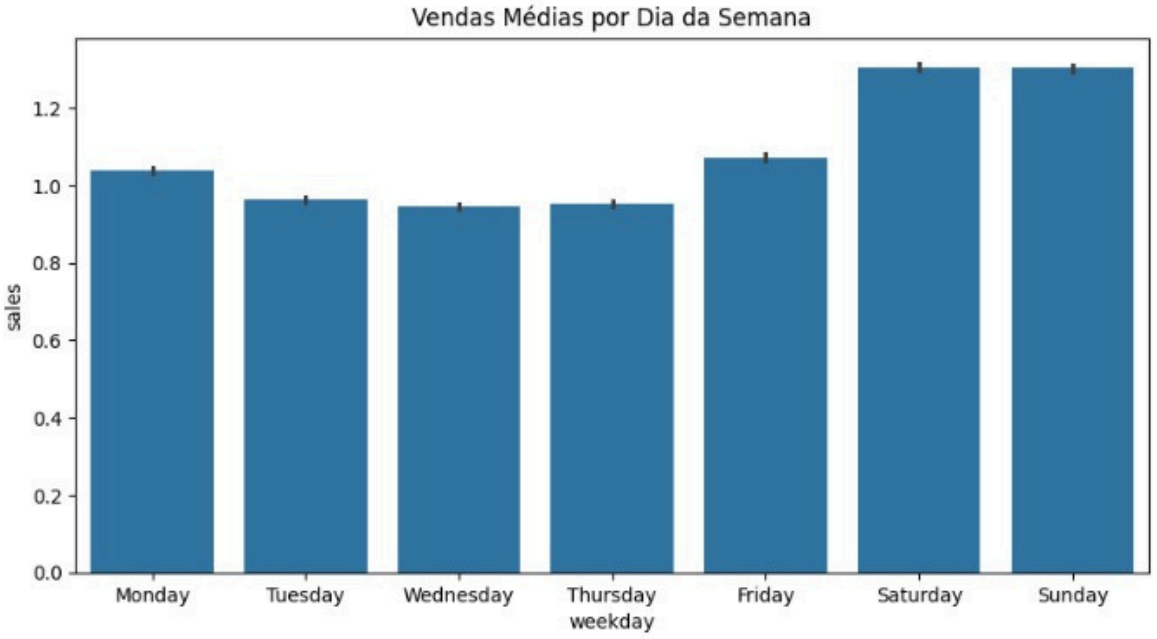
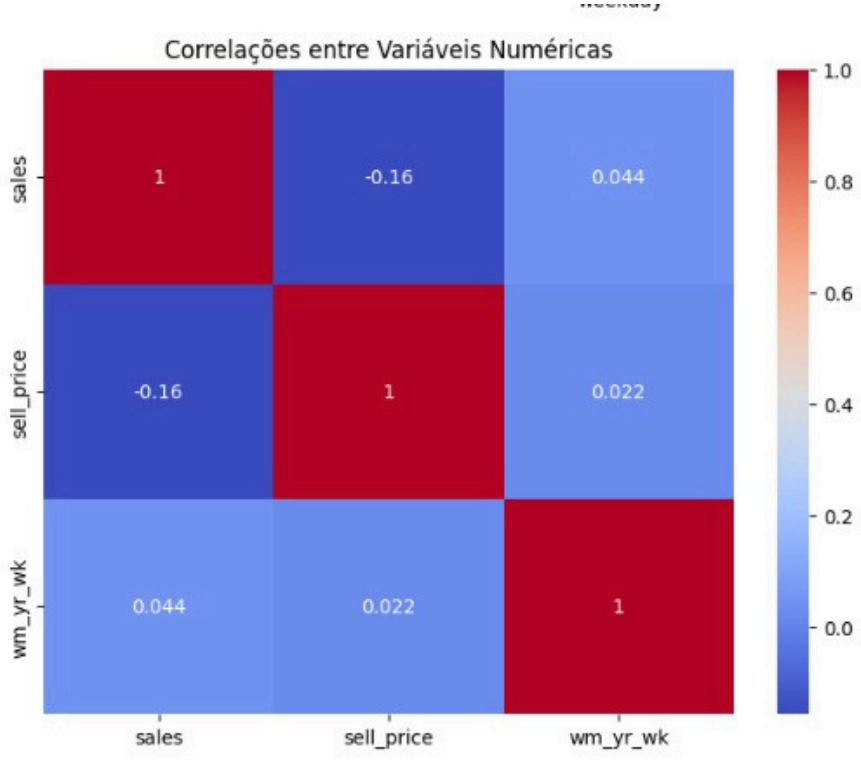


Gráfico: Correlações entre Variáveis Numéricas

- A matriz de correlação revela relações estatísticas entre as principais variáveis numéricas analisadas. A correlação entre “sell_price” e “sales” é negativa (-0,16), confirmando a tendência observada anteriormente: quanto maior o preço, menor o volume de vendas. As demais correlações são próximas de zero, indicando que não há fortes relações lineares entre as variáveis, o que sugere a necessidade de modelos mais sofisticados para captar as interações não-lineares que possam existir entre elas.



Conclusão Final — Integrada com Todos os Gráficos

A análise abrangente dos dados de vendas do Walmart, por meio de diversas visualizações gráficas e estatísticas, permitiu revelar aspectos fundamentais sobre o comportamento do mercado e as operações logísticas da rede.

Primeiramente, observou-se que embora a maioria das datas não esteja associada a eventos sazonais, determinadas ocasiões como o Super Bowl e o Valentine's Day podem oferecer oportunidades estratégicas para ações promocionais específicas. A distribuição de vendas por estado confirmou a liderança absoluta da Califórnia, enquanto Texas e Wisconsin apresentaram desempenhos mais modestos, evidenciando a necessidade de políticas regionais diferenciadas.

A relação inversa entre preço e volume de vendas, reforçada pelos gráficos de dispersão e correlação, destacou a importância de uma política de preços sensível à elasticidade, potencializando vendas com ajustes estratégicos. Complementarmente, o boxplot por categoria e a distribuição geral das vendas evidenciaram a forte dominância e variabilidade da categoria de alimentos, que, além de ser a mais vendida, apresenta maior dispersão e maior frequência de volumes elevados.

A análise do comportamento temporal das vendas demonstrou uma tendência sazonal e crescente ao longo dos anos, com vendas intensificadas nos fins de semana, reforçando a necessidade de otimização das operações logísticas e promocionais nestes períodos de maior movimento.

Por outro lado, a distribuição extremamente assimétrica das vendas sugere que a maior parte dos produtos possui baixa rotatividade, o que reforça a importância de estratégias como o gerenciamento da curva ABC e de políticas diferenciadas para itens de alta e baixa demanda.

Por fim, a análise estatística e gráfica demonstrou que embora o volume total de vendas seja expressivo, há variabilidades importantes entre produtos, categorias, regiões e momentos do tempo. Isso reforça a necessidade de modelos de previsão robustos e personalizados, capazes de capturar as nuances do comportamento de consumo, integrando não apenas variáveis internas (preços, categorias), mas também externas (eventos sazonais, indicadores econômicos).