Projeto de Análise de Vendas

Relatório de Preparação e Análise Exploratória de Dados de Vendas

Agenda: A Jornada da Análise

Geração de Dados

Criação de 300 registros sintéticos com 6 categorias para simular o conjunto de dados base.

Inconsistências e Diagnóstico

Inserção intencional de falhas (nulos, datas inválidas) para replicar desafios de dados reais e posterior detecção.

Limpeza de Dados

Estratégias de tratamento para remoção de duplicatas, correção de tipos e imputação de campos faltantes.

Análise Exploratória (EDA)

Identificação de tendências mensais, picos de venda e desempenho dos produtos de maior faturamento.

Insights & Próximos Passos

Conclusões principais, recomendações estratégicas e o futuro da gestão de dados (SQL, Dashboards).

Etapa 1: Geração de Dados Simulados

Para a análise, foram geradas **300 linhas de vendas fictícias** cobrindo o período de **01/01/2023 a 31/12/2023**.

O dataset simula seis categorias de produtos, cada uma com distribuições específicas para refletir o comportamento de vendas do mundo real:

- **Categorias:** Eletrônicos, Casa & Cozinha, Esporte, Moda, Brinquedos, Material Escolar.
- **Quantidade:** Segue uma distribuição de Poisson, ajustada para simular a sazonalidade mensal.
- **Preço:** Usa uma distribuição log-normal, variando significativamente entre as diferentes categorias.

Amostra do Conjunto de Dados Simulado

ID	Data	Produto	Categoria	Quantidade	Preço
101	2023-01-15	Fone Pro	Eletrônicos	2	249.99
102	2023-02-03	Panelas Antiaderentes	Casa & Cozinha	1	149.50
103	2023-11-20	Mochila Escolar	Material Escolar	5	45.00

Inconsistências e Duplicatas Introduzidas

Para simular a natureza imperfeita dos dados do mundo real, erros foram intencionalmente introduzidos antes da etapa de diagnóstico.

Nulos de Categoria

~17,1%

Linhas com informações de categoria de produto faltantes.

Nulos de Preço

~14,0%

Registros onde o valor do preço estava ausente ou nulo.

Datas Inválidas

~4,8%

Datas fora do intervalo de análise válido (01/01/2023 - 31/12/2023).

Valores Não-Positivos

~1,6%

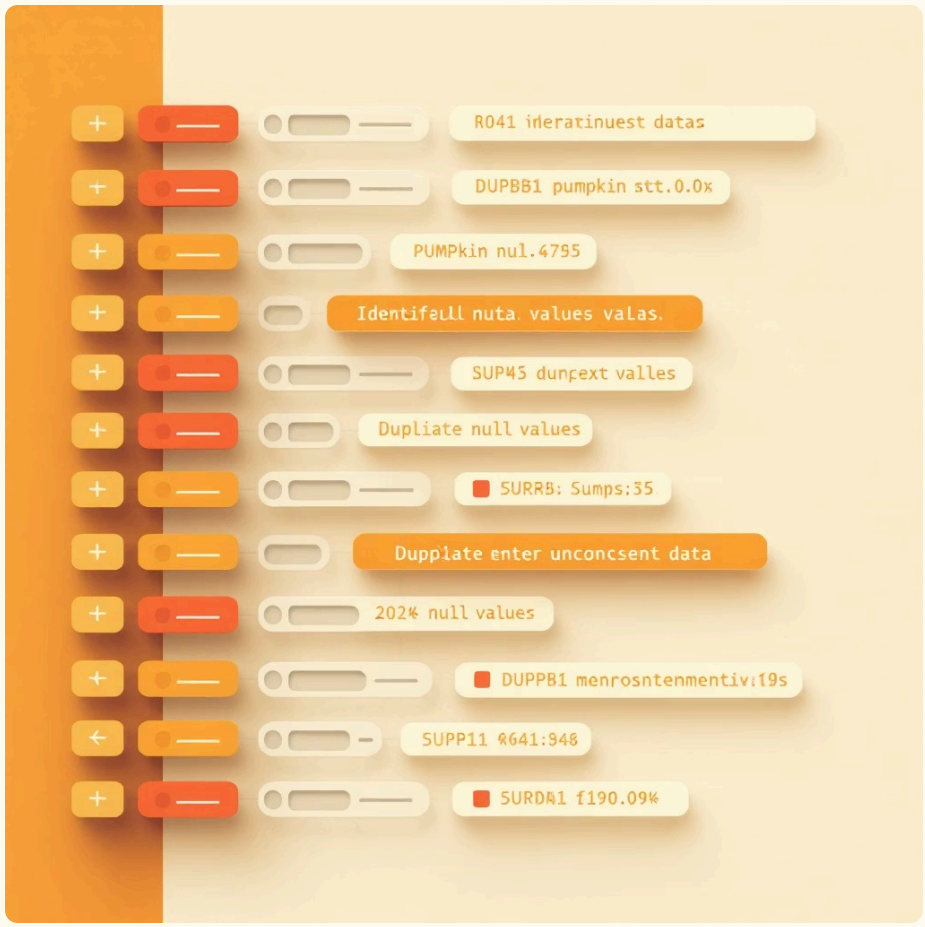
Registros com quantidade de itens igual a zero ou negativa.

Adicionalmente, 15 entradas duplicadas (baseadas no par ID e Data) foram inseridas, demandando um passo de desduplicação crítica.

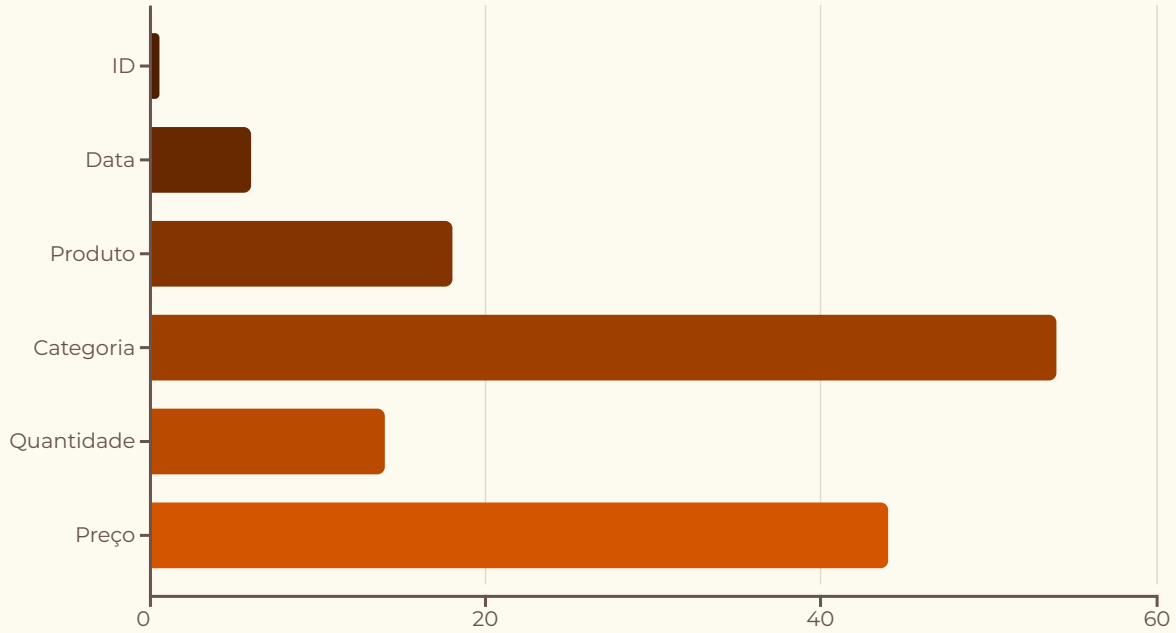
Diagnóstico do CSV (Apresentação Inicial)

A função de diagnóstico automatizado identificou a estrutura inicial da base de dados e os pontos críticos de qualidade, confirmando a necessidade de limpeza.

- **Tamanho da Base:** 315 linhas e 6 colunas.
- **Nulos:** Identificação e contagem de valores ausentes por coluna.
- **Duplicatas:** Contagem total de entradas repetidas.
- **Validação:** Detecção de datas fora do range e quantidades inconsistentes.



Contagem de Valores Nulos por Coluna



Limpeza e Transformação Essencial

A função principal de limpeza tratou as inconsistências para garantir a qualidade analítica dos dados.

Remoção de Duplicatas

Entradas repetidas foram eliminadas usando a combinação única de **(ID, Data)**.

Conversão de Tipos

Colunas **Data** (para datetime) e **Quantidade** (para inteiro) foram padronizadas.

Imputação de Nulos

Categoria preenchida pela Moda ("Moda"); Preço preenchido pela Mediana geral.

Eliminação de Inválidos

Registros com datas fora do ano de 2023 ou Quantidade ≤ 0 foram removidos.

O resultado foi o arquivo **data_clean.csv**, uma base com 291 registros, pronta para a próxima fase de análise exploratória.

Etapa 2: Carregamento e Cálculo de Totais

Com os dados limpos, a próxima fase envolveu o carregamento em um ambiente analítico (Pandas) e a criação de métricas chave para a análise de desempenho.

```
RAW_PATH  : C:\DEV\Teste_Analytics_CristovamPaulo\data\raw\vendas_2023.csv
CLEAN_PATH: C:\DEV\Teste_Analytics_CristovamPaulo\data\processed\data_clean.csv
len(df_disk) (lido do disco): 291
Resumo da limpeza:
- linhas_iniciais: 315
- datas_invalidas: 9
- duplicatas_exatas: 0
- ids_duplicados: 15
- quantidade_nao_positiva_para_NaN: 2
- preco_nao_positivo_para_NaN: 0
- preco_imputado: 42
- quantidade_imputada: 16
- duplicatas_por_chave_encontradas: 0
- produto_imputado: 2
- categoria_imputada: 0
- linhas_finais: 291
- arquivo_limpo: C:\DEV\Teste_Analytics_CristovamPaulo\data\processed\data_clean.csv
Dtypes:
ID                    int64
Data          datetime64[ns]
Produto       string[python]
Categoria     string[python]
Quantidade             Int64
Preco                float64
dtype: object
...
7     586.50
8    2050.40
9     488.80
dtype: float64
```

Transformações Aplicadas:

- Dados limpos carregados com sucesso (291 linhas).
- Conversão final do campo Data para o tipo datetime.
- Criação da coluna **Total (R\$)**: O faturamento de cada transação.

Esta métrica total é fundamental para as análises de tendências e faturamento que se seguirão.

```
=== Prévia: Top 5 (visualização) ===
```

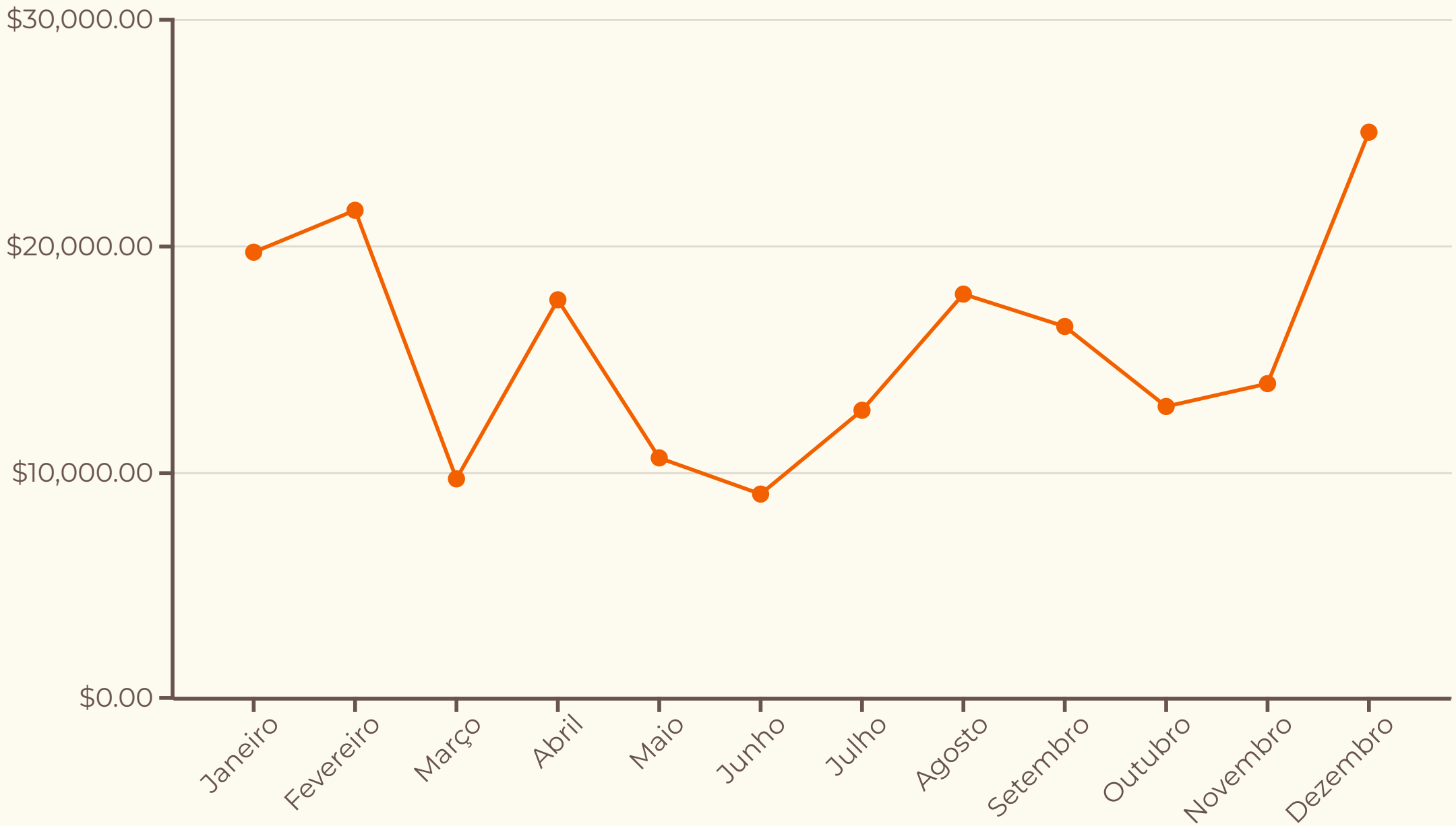
	Produto	Total
0	Caixa Bluetooth	25147.72
1	Fone Pro	16815.82
2	Smartwatch	14902.39
3	Mouse Óptico	10154.66
4	Bola Oficial	8818.55

```
Total geral de vendas: R$ 187.183,68

Produto campeão: Caixa Bluetooth com total de R$ 25.147,72
```

Tendência Mensal de Vendas (Faturamento Total)

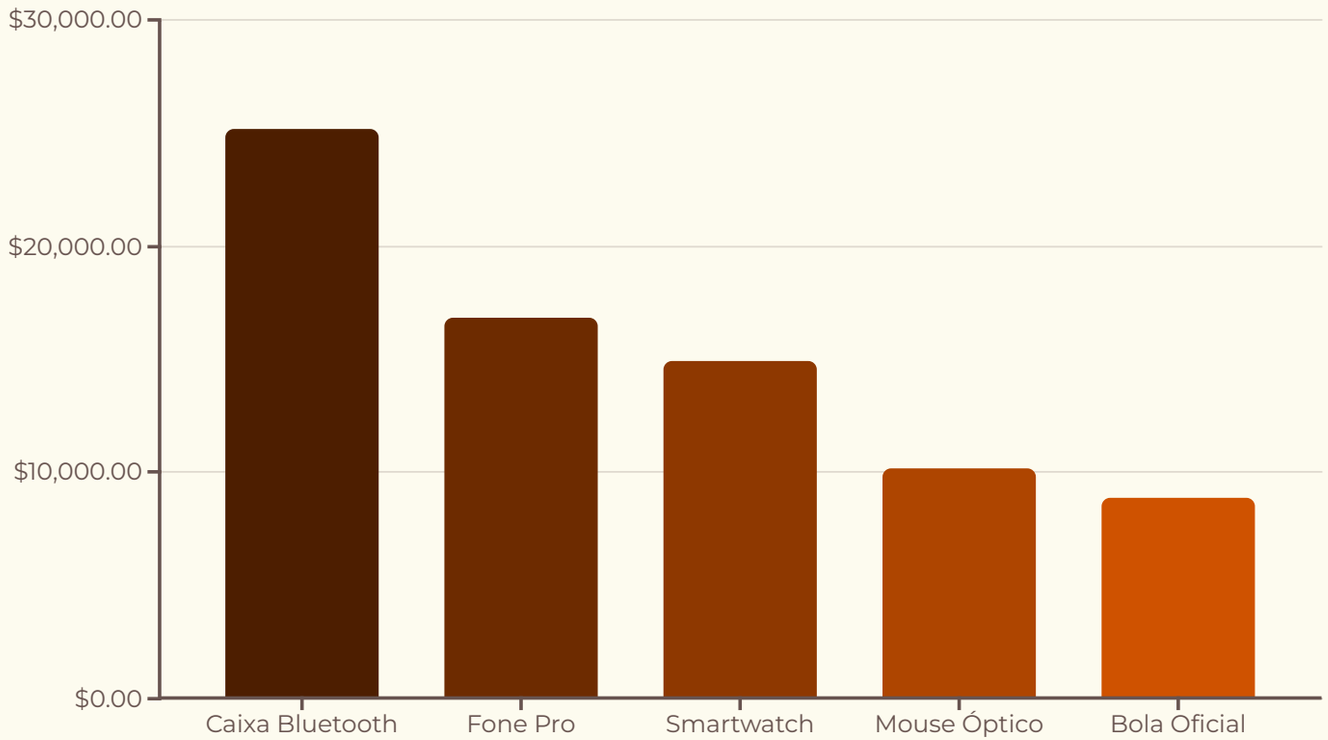
A análise do faturamento mensal revela uma sazonalidade clara, com picos estratégicos ao longo do ano.

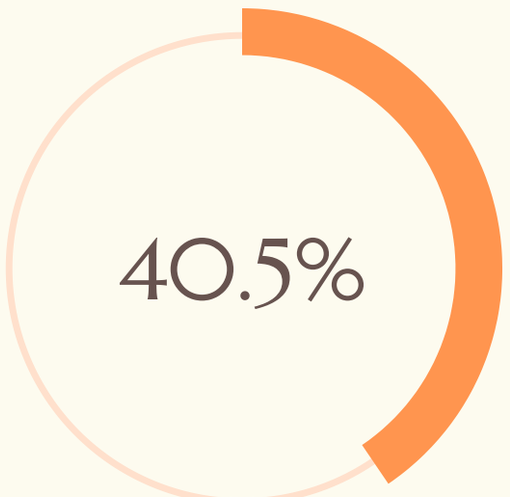


Os picos em **Janeiro**, **Fevereiro** e **Dezembro** merecem atenção especial para planejamento de estoque e marketing, enquanto junho e Maio registram as maiores quedas.

Top 5 Produtos por Faturamento Anual

A maior parte do faturamento total é concentrada em um pequeno número de produtos, evidenciando o fenômeno da "Cauda Longa".





Faturamento Concentrado

Estes 5 produtos representam quase 40,5% de toda a receita de 2023.

Quatro dos cinco produtos mais vendidos pertencem à categoria Eletrônicos, indicando alta dependência.

Insights & Recomendações Estratégicas

Picos de Venda Claros

dezembro, fevereiro e janeiro concentram 35% da receita anual, exigindo planejamento antecipado de estoque e campanhas.

Dependência de Eletrônicos

Eletrônicos gera 74,7 mil (a maior parte do faturamento).
Risco/Oportunidade de diversificação.

Queda em Jun/Mar

Os meses de junho e Março apresentam quedas acentuadas e são ideais para focar em promoções, pacotes e descontos agressivos.