🔍 3. Análise Exploratória dos Dados (EDA)

Após a coleta e o tratamento dos dados, chega o momento de **explorá-los em profundidade**. Essa fase é essencial para **entender o comportamento das variáveis**, **descobrir padrões**, **detectar relações escondidas** e **obter os primeiros insights** que orientarão a construção dos modelos preditivos.

📊 Estatísticas Descritivas

As estatísticas descritivas são medidas que **resumem o comportamento dos dados**.
Entre as principais, destacam-se:

- **Média:** valor médio das observações (tendência central).
- **Mediana:** valor central, menos sensível a valores extremos (outliers).
- **Desvio Padrão:** mede o quanto os dados se afastam da média.
  - Um desvio padrão alto indica grande variação.
  - Um desvio padrão baixo indica dados mais estáveis.

📘 *Exemplo:*
Se a média semanal de vendas é **1.000 unidades**, mas o desvio padrão é **300**, isso indica que há **variação significativa**, podendo estar ligada a fatores como clima, feriados ou promoções.

✅ 1. Setup e Carregamento de Dados

Antes de gerar estatísticas, é preciso garantir que os dados estão prontos para análise:

Tipo de célula	Função	Código-chave
`# %% [setup] Variáveis e imports base`	Prepara o ambiente e importa bibliotecas (`pandas` , `numpy` , `matplotlib`).	`import pandas as pd` , `pd.set_option(...)`
`# %% [data] Carregamento de dados (CSV ou Parquet)`	Lê o arquivo indicado em `DATA_PATH` e mostra as primeiras linhas (`df.head(3)`).	`pd.read_csv(DATA_PATH)` ou `pd.read_parquet(DATA_PATH)`

Essas duas células garantem que o **dataset está limpo e carregado na variável** `df` — pré-requisito para todas as análises seguintes.

✅ 2. Estatísticas Descritivas Ampliadas

Esta é a célula **central da etapa Estatísticas Descritivas**, e corresponde exatamente ao trecho que você descreveu.

Tipo de célula	Bloco	O que faz
`# %% [eda] Estatísticas descritivas ampliadas`	Calcula e exibe **resumos estatísticos** de todas as variáveis numéricas.	

Dentro dela:

- `df.describe().T` → mostra média, mediana (via `.50%`), mínimo, máximo, desvio padrão (`std`);
- `df.skew()` e `df.kurtosis()` → medem **assimetria** e **achatamento** da distribuição;
- `df.isna().sum()` e `df.duplicated().sum()` → verificam **valores ausentes e duplicados**;
- Mostra também **número de valores únicos e tipo de dado** (`dtype`).

💡 **É aqui que você vai ver na prática:**

- Quais colunas têm maior variação (desvio padrão alto);
- Quais são mais estáveis (desvio padrão baixo);
- Se existem outliers ou dados incompletos.

🔗 Correlações entre Variáveis

A correlação mede **o grau de relacionamento entre duas variáveis**.
O valor vai de **-1 a +1**:

- **+1:** relação positiva perfeita (quando uma sobe, a outra também sobe).
- **-1:** relação negativa perfeita (quando uma sobe, a outra desce).
- **0:** sem relação linear significativa.

📈 *Exemplo prático:*
No nosso caso, observamos **correlação negativa entre temperatura e volume de vendas (-0.22)**.
Isso significa que, **em períodos mais quentes, as vendas tendem a cair**, indicando **um**

comportamento sazonal — possivelmente os consumidores compram menos certos produtos no calor.

⚖️ Distribuição de Classes

Quando o problema envolve **classificação** (por exemplo: “comprou” ou “não comprou”, “alto risco” ou “baixo risco”), é importante analisar **a proporção de cada classe**.
Um **desequilíbrio muito grande** pode prejudicar o modelo de previsão.

📘 *Exemplo:*
Se 90% das vendas são de um único tipo de produto, o modelo pode se tornar “viciado” e errar previsões das demais categorias.

🧭 Visualizações e Insights Gráficos

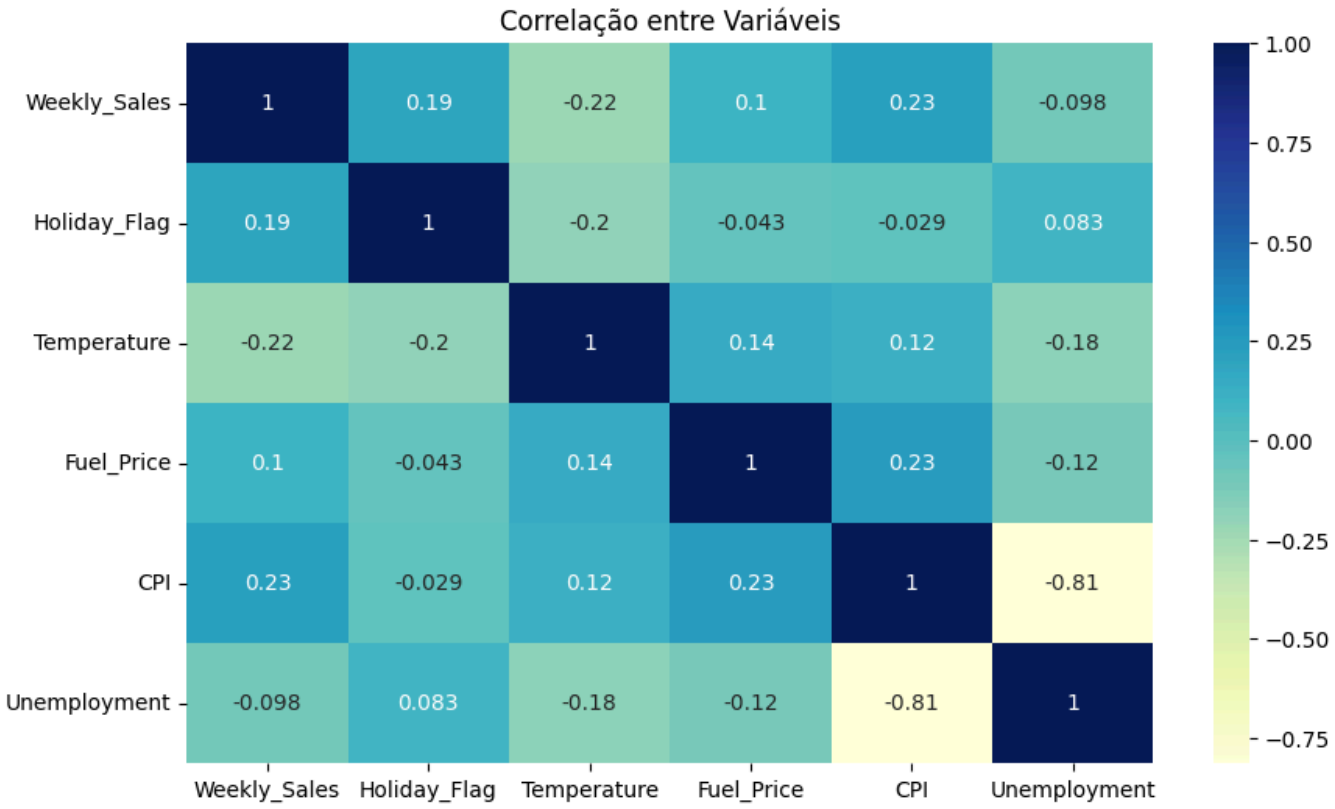
Os gráficos ajudam a **traduzir números em histórias visuais**, tornando mais fácil a compreensão por parte de gestores e equipes não técnicas.

Alguns exemplos de visualizações aplicadas:

- **Histogramas:** mostram a distribuição de uma variável (ex.: faixas de preço ou temperatura).
- **Boxplots:** evidenciam outliers e dispersão.
- **Heatmaps de correlação:** destacam variáveis fortemente relacionadas.
- **Gráficos de dispersão (scatterplots):** mostram relações entre duas variáveis.

📊 *Insight gerado:*
“Identificamos correlação negativa entre temperatura e volume de vendas (-0.22), indicando efeito sazonal e necessidade de ajustar promoções em períodos de calor.”

Correlação entre Variaveis



Analisando Sazonalidade.

📈 Tendência e Sazonalidade nas Vendas Semanais

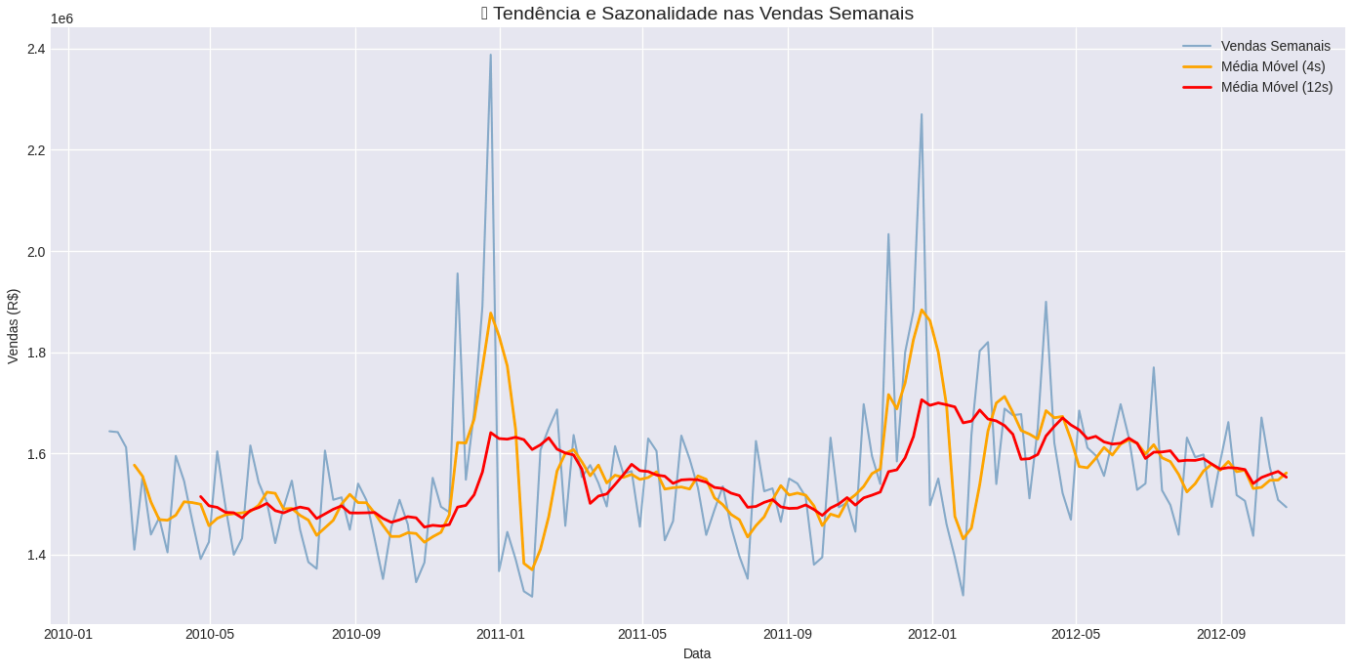
🧩 1️⃣ Tendência e Sazonalidade nas Vendas Semanais

📊 **Gráfico:** linha azul (vendas semanais), médias móveis (4 e 12 semanas).

Explicação para público leigo:

> Este gráfico mostra a variação natural das vendas semanais ao longo do tempo.
>
> As linhas coloridas representam **médias móveis**, que suavizam as oscilações e ajudam a visualizar tendências de curto e médio prazo:
>
> - A **média móvel de 4 semanas** (linha laranja) mostra o comportamento mais imediato das vendas.
> - A **média móvel de 12 semanas** (linha vermelha) revela tendências mais estáveis e sazonais.
>
> Observa-se que existem períodos com **altas e baixas recorrentes**, indicando uma **sazonalidade natural** — períodos do ano onde a demanda cresce (ex: datas comemorativas) e períodos de retração.

📌 *Mensagem para o board:*
A empresa deve acompanhar essas tendências para planejar **estoques e promoções** de acordo com os picos e quedas sazonais.



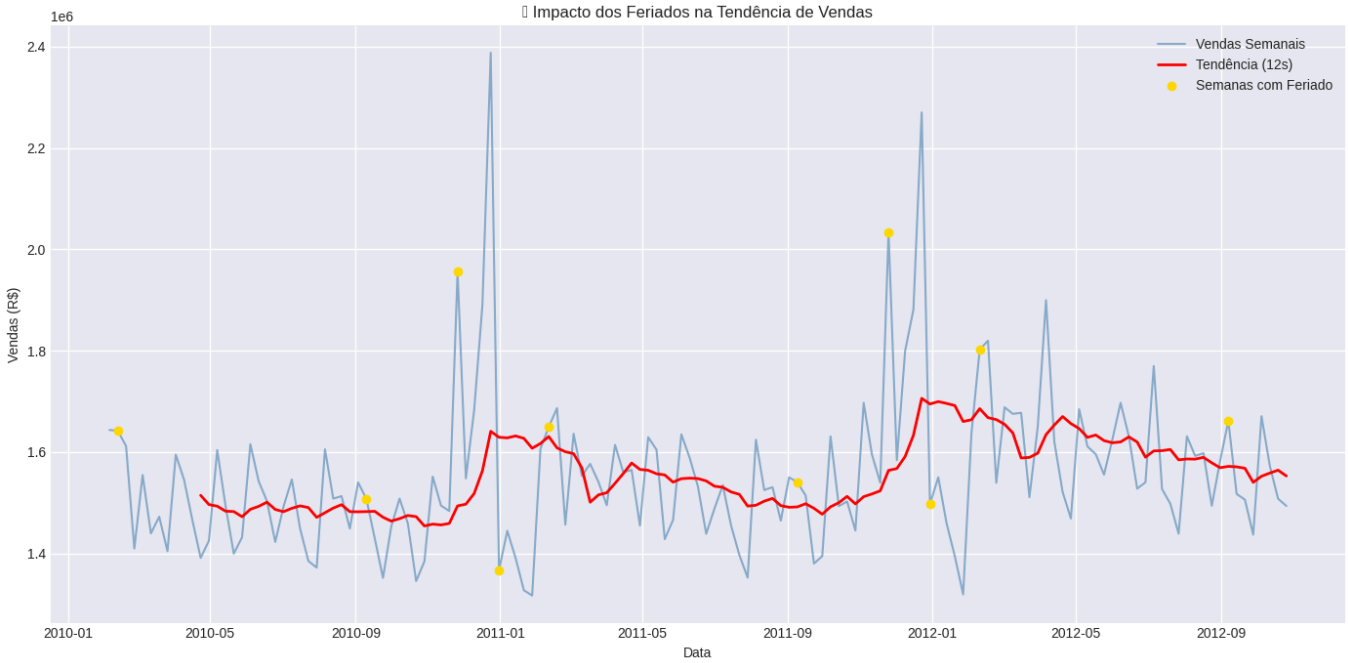
🧩 2️⃣ Impacto dos Feriados na Tendência de Vendas

📊 **Gráfico:** mesma linha de tendência (12s), com **marcação de feriados** em pontos amarelos.

Explicação executiva:

> Aqui mostramos o impacto dos **feriados** sobre o comportamento das vendas.
>
> Os pontos amarelos representam semanas que coincidiram com feriados.
> É possível notar que, em boa parte dos casos, **as vendas aumentam antes ou durante feriados** — reflexo do maior consumo e reposição de estoques.
>
> Após esses períodos, ocorre uma **pequena queda natural**, à medida que o consumo se normaliza.

📌 *Mensagem para o board:*
Essas evidências ajudam a **prever picos de demanda** e **ajustar campanhas** de marketing e logística em datas específicas do calendário.



🧩 3️⃣ Evolução de Vendas e Ciclos Sazonais

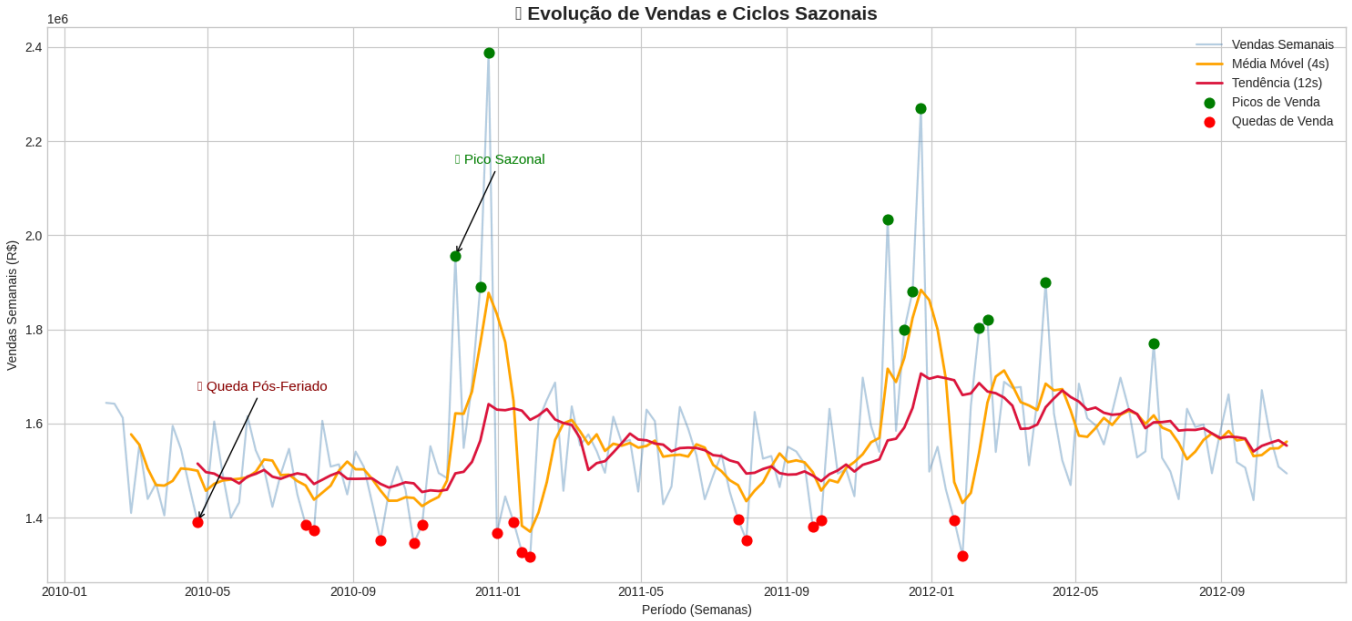
📊 **Gráfico:** vendas semanais + médias móveis + **pontos de pico (verde)** e **queda (vermelho)** com anotações visuais.

Explicação (storytelling):

> Este gráfico combina todos os elementos anteriores e mostra a **história completa do comportamento das vendas**.
>
> - Os **picos sazonais (verdes)** indicam momentos de forte aumento, geralmente associados a promoções, feriados ou condições climáticas favoráveis.
> - As **quedas pós-feriado (vermelhas)** evidenciam períodos de retração, em que o consumo desacelera.
>
> A leitura visual é clara: o ciclo se repete em ondas de crescimento e queda.
> Esse padrão é típico de mercados com forte componente **sazonal e emocional de compra**.

📌 *Mensagem para o board:*
Com base nesses padrões, a empresa pode:

- **Antecipar ações promocionais** antes dos picos previstos;
- **Evitar excesso de estoque** após períodos de alta;
- **Planejar campanhas climáticas e sazonais** (ex: calor, feriados, datas regionais).



🎯 Conclusão da Etapa

A **Análise Exploratória dos Dados (EDA)** é onde o cientista de dados **descobre o "DNA" do conjunto de dados**.
Essa compreensão é a base para:

- escolher os **melhores algoritmos** de previsão;
- ajustar variáveis que realmente impactam o negócio;
- e **comunicar com clareza** ao time executivo quais fatores são críticos para o desempenho da empresa.