Distribuição Normal da variável dependente: Importância e Aplicação

🎯 Pergunta:

Testar se o consumo de cerveja se distribui como uma normal é útil? Como isso influencia o modelo e como podemos usar essa informação?

☑ Sim, é útil!

E aqui está o **porquê**:

Razão	Explicação
1. Validação de hipóteses estatísticas	Muitos testes estatísticos clássicos (como testes t, construção de intervalos de confiança para a média, análise de resíduos) assumem que os dados seguem uma distribuição normal.
2. Melhor interpretação do erro	Se o erro (resíduo) entre o valor real e o predito for normalmente distribuído, conseguimos construir intervalos de confiança para previsões futuras e fazer inferência estatística correta.
3. Aplicabilidade de técnicas paramétricas	Se o consumo seguir uma normal, podemos usar modelos paramétricos com maior confiança (como regressão linear simples, métodos baseados em inferência gaussiana).
4. Previsibilidade e Simplicidade	Dados que seguem normalidade tendem a ter comportamento mais estável em modelagem, reduzindo necessidade de transformações complexas.

❌ Se não for normal:

Se você descobrir que **o consumo de cerveja não é normalmente distribuído**, isso **não inviabiliza** o modelo, mas te alerta para:

Possíveis **outliers** ou **assimetria** (que podem ser corrigidos ou tratados).
A necessidade de considerar modelos **não paramétricos** (ex: regressão quantílica, árvores de decisão).
Possível necessidade de **transformações nos dados** (como `log()`, `sqrt()`, Box-Cox) para normalizar os resíduos.
Avaliar se a violação de normalidade afeta **apenas o Y** (consumo) ou os **resíduos do modelo**.
➔ **O mais importante na regressão linear é a normalidade dos resíduos**, não necessariamente da variável dependente!

📌 Como usar essa informação na prática

Situação	Ações práticas
Dados seguem normal	Pode seguir com regressão linear, testes t, construção de intervalos, previsão baseada em erro padrão.
Dados não seguem normal	Analisar outliers, fazer transformações, considerar robustez dos métodos (regressões robustas, árvores, etc.).
Resíduos são normais, mesmo que Y não seja	Pode usar o modelo normalmente. O foco principal é que os **resíduos** sejam aproximadamente normais para garantir boa inferência.

🛠️ Ferramentas para Testar Normalidade

`scipy.stats.normaltest()` (teste de D'Agostino e Pearson).
`scipy.stats.shapiro()` (teste de Shapiro-Wilk).
`statsmodels.qqplot()` (gráfico Q-Q para avaliar visualmente).

🎯 Conclusão

✅ **Testar a normalidade** ajuda a saber **como ajustar melhor o modelo** e a entender o **comportamento dos erros**.

✅ **Normalidade dos resíduos** é mais importante do que da variável dependente para a regressão linear.

✅ Mesmo que os dados não sejam normais, é possível trabalhar com modelos **alternativos ou transformados**.