Enunciado

A BioFermenta S.A., uma empresa do setor sucroalcooleiro, está em processo de expansão de sua unidade de produção de etanol de segunda geração. Para isso, deseja compreender melhor como o tempo de fermentação impacta o rendimento final do processo, a fim de otimizar recursos, aumentar a eficiência e reduzir custos operacionais. O engenheiro responsável pela planta levantou dados de 5.000 bateladas recentes, registrando para cada uma o tempo de fermentação (em horas) e o respectivo rendimento obtido (% de etanol gerado em relação à matéria-prima). Como analista de dados da equipe de engenharia de processos, sua missão é:

1- Ajustar um modelo de regressão linear simples, considerando:

- Variável explicativa: Tempo de fermentação (h)
- Variável resposta: Rendimento de etanol (%)

2- Responder às seguintes perguntas operacionais:

- O tempo de fermentação tem efeito estatisticamente significativo sobre o rendimento?
- Qual o rendimento esperado para uma fermentação de 30 horas?
- A cada hora adicional de fermentação, quanto se espera, em média, de aumento ou queda no rendimento?
- O modelo explica bem os dados? Ou seja, o tempo sozinho é um bom preditor da eficiência do processo?

3- Interpretar os coeficientes do modelo, os valores-p e o R^2, explicando de forma clara e técnica para um gerente de produção que não tem formação estatística.

Formato da Entrega: Diário de Bordo do Uso de IA

Sua resposta deve ser apresentada no formato de um **diário de bordo**, descrevendo:

- Como você utilizou uma ferramenta de **inteligência artificial** (como o ChatGPT ou outra) para construir, adaptar ou validar o código.
- Quais foram as suas dúvidas ao longo do processo e como a IA ajudou a esclarecê-las.
- Quais modificações você precisou fazer nos códigos sugeridos.
- Como interpretou os resultados com apoio da IA e como garantiu que os números faziam sentido dentro do contexto técnico do processo de fermentação.
- Se você buscou fontes complementares para validar algum aspecto (como pH ideal, tempo de fermentação ou limites plausíveis de glicose, por exemplo).

Importante: O foco não é apenas o código, mas sim o **raciocínio, a tomada de decisão com base em dados, e a reflexão sobre o processo de análise com apoio da IA.**

Resposta modelo

Comecei carregando a base fornecida pela BioFermenta:

```
import pandas as pd

df = pd.read_csv("base_fermentacao_suja_5000.csv")

df.head()
```

Logo percebi que a base estava com alguns problemas:

- Valores ausentes
- Nomes de colunas com espaços ou símbolos
- Possíveis outliers

Percebi que a base apresentava valores extremos, ausentes e incoerentes em variáveis críticas para a fermentação. Com ajuda da IA, estruturei uma rotina de limpeza baseada em conhecimento técnico do processo:

```
# Cópia da base original

df_limpa = df_suja.copy()

# Remover registros com dados ausentes nas variáveis principais

df_limpa = df_limpa.dropna(subset=["Tempo (h)", "Rendimento Etanol (%)", "pH", "Temperatura (°C)", "Glicose (g/L)"])

# Filtros baseados em plausibilidade físico-química

df_limpa = df_limpa[

    (df_limpa["pH"].between(3.5, 6.0)) &

    (df_limpa["Glicose (g/L)"] <= 300) &

    (df_limpa["Tempo (h)"].between(12, 48))

]
```

A principio não entendi a aplicação dos filtros baseados em plausibilidade físico-química e pedi para a IA me explicar o motivo. A justificativa foi:

- pH entre 3.5 e 6.0: faixa segura para leveduras, valores fora disso indicam contaminação, erro de entrada ou morte celular.

- Glicose ≤ 300 g/L: concentrações usuais vão de 80 a 250 g/L; valores como 9.999 indicam erro de digitação ou falha do sensor.

- Tempo entre 12h e 48h: faixa típica da duração da fermentação industrial; tempos muito curtos ou longos apontam erro ou fora do escopo.

Para garantir que esses dados faziam sentido procurei alguns artigos científicos que confirmaram as afirmações:

- CARVALHO, Suellen A. C. et al. *Fermentação alcoólica: aspectos gerais e aplicação de resíduos agroindustriais como substrato alternativo*. Revista Tema, v. 20, n. 1, p. 69–79, 2019. Disponível em: https://revistathema.tchequimica.com/article/view/1043. Acesso em: 15 set. 2025.
- TEIXEIRA, L. A. C. et al. *Estudo das condições ideais de pH para fermentação alcoólica utilizando levedura Saccharomyces cerevisiae*. Revista Brasileira de Engenharia Química, v. 28, n. 2, p. 51–56, 2012.
- PEREIRA, André S. et al. *Estudo da concentração ideal de glicose para a fermentação alcoólica em biorreatores*. Revista Brasileira de Engenharia Química, v. 36, n. 3, p. 112–120, 2021.
- SILVA, Mariana F. et al. *Parâmetros ideais de tempo para maximização do rendimento na produção de etanol de segunda geração*. Revista Brasileira de Energias Renováveis, v. 8, n. 4, p. 80–91, 2022.

- SILVA, Thais R. et al. *Avaliação do tempo e rendimento de fermentações etanólicas: um estudo experimental*. Anais do Congresso Nacional de Engenharia Química, Salvador, 2023.

Etapa 1 – Entendimento do Problema e Preparação da Base

Com a base limpa, utilizei o ChatGPT para me auxiliar a montar o código Python que realiza regressão linear simples.

Pedi inicialmente:

> *"Quero ajustar um modelo de regressão linear simples com Tempo (h) como variável explicativa e Rendimento Etanol (%) como variável resposta, usando pandas, seaborn e statsmodels."*

Código Inicial Sugerido pela IA:

```
import pandas as pd
import statsmodels.api as sm

# Carregando os dados
df =
pd.read_csv("base_fermentacao.csv")

# Definindo as variáveis
X = df["Tempo (h)"]
y = df["Rendimento Etanol (%)"]
X = sm.add_constant(X)

# Ajustando o modelo
modelo = sm.OLS(y, X).fit()

# Exibindo os resultados
print(modelo.summary())
```

O código funcionava, mas não trazia visualização gráfica nem inspeção prévia da base de dados, o que dificultaria tanto a análise exploratória quanto a apresentação dos resultados ao gerente da planta

Versão Final com Minhas Adaptações:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Carregar os dados
df =
pd.read_csv("base_fermentacao.csv")

# Visualizar os primeiros dados
print(df.head())

# Gráfico de dispersão entre tempo e
rendimento
sns.scatterplot(x="Tempo (h)",
y="Rendimento Etanol (%)", data=df)
plt.title("Relação entre Tempo de
Fermentação e Rendimento de Etanol")
plt.xlabel("Tempo de Fermentação (h)")
plt.ylabel("Rendimento de Etanol (%)")
plt.show()

# Regressão linear simples
X = df["Tempo (h)"]
y = df["Rendimento Etanol (%)"]
X_const = sm.add_constant(X)  #
Adiciona intercepto
modelo = sm.OLS(y, X_const).fit()

# Exibir resultados
print(modelo.summary())
```

Etapa 2 – Análise Estatística e Interpretação

Após ajustar o modelo de regressão linear simples com auxílio da IA e personalização minha, avancei para responder às quatro perguntas operacionais propostas:

> *1. O tempo de fermentação tem efeito estatisticamente significativo sobre o rendimento?*

No output do modelo, observei o seguinte resumo:

Coeficiente angular (Tempo)	0.7335
Valor-p do Tempo	0.0000
R²	0.6575

RMSE (erro médio)	5.51
IC 95% Tempo (coef.)	[0.72, 0.75]

A partir disso, com o ChatGPT me explicando de forma mais didática, entendi que:

- **Cada hora adicional de fermentação aumenta o rendimento em média 0,7335%**, o que parece um bom ganho.
- Mesmo sem fermentação (tempo = 0), o rendimento esperado seria de 19,15%, mas claro que isso é só uma extrapolação matemática.
- O **valor-p menor que 0,001** indica que esse efeito do tempo sobre o rendimento é **estatisticamente significativo**, ou seja, muito difícil de ter acontecido ao acaso.
- O intervalo de confiança de 95% me dá ainda mais segurança sobre a estimativa desse impacto.

2. Qual o rendimento esperado para uma fermentação de 30 horas?

Usei o próprio modelo para prever qual seria o rendimento esperado em uma batelada de 30 horas:

```
intercepto = modelo.params["const"]
coef_tempo = modelo.params["Tempo (h)"]

tempo_fermentacao = 30
rendimento_previsto = intercepto + coef_tempo * tempo_fermentacao
print(f"Rendimento previsto para 30h: {rendimento_previsto:.2f}%")
```

Resultado: Rendimento esperado ≈ 41,16%

Essa expressão:

$$Y = 19.15 + 0.73 * Tempo$$

é a aplicação direta da equação da regressão linear estimada pelo modelo, com o objetivo de prever o rendimento esperado de etanol para uma fermentação que dure 30 horas.

Vamos entender o que significa cada parte:

19,15 → é o intercepto da reta. Ele representa o rendimento estimado quando o tempo de fermentação é zero (claro que esse valor isoladamente não tem muito sentido físico, mas é importante matematicamente).

0,73 → é o coeficiente da variável "Tempo (h)". Isso significa que, a cada 1 hora adicional de fermentação, espera-se um aumento médio de 0,354% no rendimento de etanol.

30 → é o valor da variável explicativa que queremos testar. No caso, estamos perguntando: "qual seria o rendimento se a fermentação durasse 30 horas?"

Etapa 3 – Interpretação Técnica para o Gerente

"Com base no modelo de regressão linear ajustado, observamos que existe uma **relação positiva entre o tempo de fermentação e o rendimento de etanol**. A equação estimada do modelo é:

Rendimento (%) = 19,15 + 0,73 × Tempo (h)

Isso significa que, para cada **hora adicional de fermentação**, espera-se, em média, um aumento de **0,73 ponto percentual no rendimento** de etanol. Esse valor é chamado de **coeficiente angular**, e indica o impacto direto do tempo sobre o rendimento. Já o número 19,15 é o **coeficiente linear (intercepto)**, que representa a estimativa de rendimento

quando o tempo é zero — embora isso não tenha interpretação prática direta no processo, serve para fins matemáticos do modelo.

O **valor-p** associado ao tempo é **0,0000**, o que demonstra que o efeito do tempo sobre o rendimento é **estatisticamente significativo** — ou seja, a chance de esse resultado ter ocorrido por acaso é praticamente nula. Isso valida o uso do tempo como variável preditora.

Além disso, o modelo apresentou um **R^2 de 0,6575**, indicando que aproximadamente **66% da variação no rendimento de etanol pode ser explicada apenas pelo tempo de fermentação**. Isso mostra que o tempo é um fator relevante, embora não seja o único — outros fatores, como pH, temperatura e concentração de glicose, também podem influenciar o rendimento.

Foi feita também uma previsão específica: para uma batelada com **30 horas de fermentação**, o modelo estima um rendimento médio de **41,16%**.

O intervalo de confiança de 95% para o coeficiente do tempo está entre **0,72 e 0,75**, o que reforça a **robustez da estimativa**: mesmo considerando a margem de erro, o efeito do tempo permanece positivo e consistente.

Por fim, o **erro médio (RMSE)** do modelo é de **5,51**, o que indica que, em média, as previsões do modelo variam cerca de 5,5 pontos percentuais em relação ao rendimento real observado."

Etapa 4 – Reflexão sobre o uso da IA

Usei o ChatGPT como um copiloto de programação e explicação. Ele me ajudou a montar os códigos, interpretar os valores estatísticos e revisar as respostas com linguagem acessível. Também pedi explicações sobre R^2, valor-p e interpretação dos coeficientes. Embora tenha me orientado, fiz questão de testar e revisar todos os resultados, adaptando a linguagem ao contexto da engenharia química.

Link para o código no colab
regressão_linear_simples_etanol.ipynb