# Universidade Positivo

Francisco Bueno Ghizelini  
 Guilherme Candida

Relatório Geral de Atividades – Projeto "Preço Justo PR

Curitiba, Paraná

2025

Francisco Bueno Ghizelini  
 Guilherme Candida

Análise de Dados do Projeto *Preço Justo PR*

Relatório do trabalho A2 da disciplina de

Data science, apresentada pelo professor Rubem Matimoto Koide pela Universidade Positivo

Curitiba, Paraná

2025

# 

**Data:** 27 de outubro de 2025

**Objetivo do Projeto:** Realizar uma análise de regressão sobre dados de preços de combustíveis no estado do Paraná, com foco no comportamento do preço da gasolina ao longo do tempo. As atividades incluíram a implementação de modelos de regressão linear e não linear, a resolução de desafios técnicos e a criação de uma ferramenta de visualização interativa (dashboard).

## **1. Preparação e Tratamento dos Dados**

Os dados foram extraídos do conjunto oficial de preços de combustíveis do programa *Preço Justo PR*.  
 As etapas de preparação incluíram:

* **Leitura do arquivo Excel** no ambiente Google Colab com a função pd.read\_excel().
* **Conversão de datas** para o formato temporal (datetime) e dos valores de venda para numérico (float).
* **Filtragem dos dados** apenas para o produto “GASOLINA”.
* **Agrupamento por data** para cálculo do preço médio diário.
* Criação da variável **“Dias”**, representando o número de dias desde a primeira data registrada, utilizada como variável independente nos modelos.

Um gráfico inicial de dispersão foi gerado para visualizar o comportamento geral dos preços ao longo do tempo.

## **2. Modelagem de Regressão**

Foram implementados e comparados cinco tipos de modelagem:

### **2.1 Regressão Linear Simples**

O modelo y=a⋅x+by = a \cdot x + by=a⋅x+b foi ajustado utilizando o método de **Mínimos Quadrados Ordinários (OLS)**, fornecido pelo LinearRegression da biblioteca *scikit-learn*.  
 Esse modelo capturou a tendência linear média de aumento dos preços ao longo do tempo.

### **2.2 Regressão Polinomial (Grau 2 – Parábola)**

O modelo y=a⋅x2+b⋅x+cy = a \cdot x^2 + b \cdot x + cy=a⋅x2+b⋅x+c foi implementado com o intuito de representar possíveis acelerações ou desacelerações nos preços.  
 O método de ajuste também foi **Mínimos Quadrados**, aplicado sobre as variáveis transformadas por PolynomialFeatures.  
 Os resultados apresentaram melhor desempenho que o modelo linear, indicando leve curvatura na tendência dos dados.

### **2.3 Regressão Não Linear (Exponencial)**

O modelo não linear da forma y=a⋅eb⋅x+cy = a \cdot e^{b \cdot x} + cy=a⋅eb⋅x+c foi ajustado a partir de diferentes métodos de otimização:

| **Método** | **Descrição** | **Biblioteca** | **Tipo** |
| --- | --- | --- | --- |
| **Levenberg–Marquardt (LM)** | Combinação entre Gauss–Newton e Gradiente Descendente. Alta estabilidade. | curve\_fit(method='lm') | Determinístico |
| **Gauss–Newton (TRF)** | Versão iterativa baseada em mínimos quadrados. | curve\_fit(method='trf') | Determinístico |
| **Máxima Verossimilhança (MLE)** | Maximiza a probabilidade dos dados observados sob hipótese de erros normais. | scipy.optimize.minimize() | Estatístico |
| **Bayesiano (MCMC)** | Baseado em inferência probabilística via amostragem de Monte Carlo. | pymc | Probabilístico |

Cada método foi avaliado por duas métricas:

* **R² (Coeficiente de Determinação):** mede o quanto o modelo explica a variância dos dados.
* **RMSE (Raiz do Erro Quadrático Médio):** indica o erro médio em unidades de preço (R$).

Gráficos foram gerados sobrepondo as curvas ajustadas de cada método aos dados reais para comparação visual.

## **3. Avaliação e Comparação dos Resultados**

Uma tabela comparativa foi gerada automaticamente, classificando os modelos pelo valor de **R² (quanto mais próximo de 1, melhor)** e **RMSE (quanto menor, melhor)**.  
 O código também identifica o **melhor modelo** de forma automática.

Os resultados indicaram que:

* O modelo **exponencial ajustado pelo método de Levenberg–Marquardt** apresentou o melhor desempenho geral, alcançando os maiores valores de R² e os menores de RMSE.
* O método **Bayesiano** obteve desempenho semelhante, com a vantagem de fornecer **intervalos de incerteza** sobre os parâmetros.
* O modelo **polinomial** também apresentou boa aderência, mas inferior aos métodos exponenciais.
* O **MLE** mostrou resultados próximos aos mínimos quadrados, confirmando a equivalência teórica sob distribuição normal dos erros.

## **4. Considerações Técnicas e Gráficas**

* Foram utilizados gráficos de dispersão e curvas ajustadas para ilustrar visualmente cada modelo.
* O projeto incluiu uma **tabela de resultados formatada com gradientes de cor** para facilitar a interpretação no Colab.
* Todas as etapas foram automatizadas, desde a leitura dos dados até a exportação dos resultados em CSV.

## **5. Conclusão Geral**

O estudo demonstrou a eficácia da **regressão não linear exponencial** na previsão e explicação do comportamento temporal dos preços de combustíveis.  
 Os métodos **Levenberg–Marquardt** e **Bayesiano (MCMC)** destacaram-se por oferecerem alta precisão e estabilidade nos ajustes.

A análise também evidenciou a importância de comparar diferentes métodos de estimação para garantir resultados confiáveis, especialmente em fenômenos econômicos que apresentam **comportamento não linear**.

O projeto cumpriu integralmente os objetivos propostos, fornecendo um **código robusto, comparativo e visualmente interpretável**, aplicável em estudos reais de séries temporais.

## **6. Sugestões de Trabalhos Futuros**

* Testar modelos com variáveis externas (ex.: inflação, preço do petróleo, câmbio).
* Explorar regressões logísticas ou modelos ARIMA.
* Expandir a abordagem Bayesiana para inferência preditiva com intervalos de confiança.
* Desenvolver uma interface interativa completa com *Streamlit* ou *Dash*.

## **7. Bibliotecas Utilizadas**

* **Pandas**, **NumPy**, **Matplotlib**, **SciPy**, **scikit-learn**, **PyMC**.
* Ambiente: **Google Colab**.
* Linguagem: **Python 3.10**.

**Conclusão Geral:**

As atividades foram concluídas com sucesso. O projeto evoluiu de um script de análise estática para um dashboard interativo e robusto, capaz de carregar os dados de forma confiável e apresentar os resultados de múltiplos modelos de regressão de maneira clara e dinâmica. A jornada incluiu a resolução de problemas comuns de manipulação de dados em ambientes de nuvem, resultando em uma ferramenta final funcional e informativa para a análise dos preços de combustíveis.