Prodigy ML Internship – Task 1
Linear Regression for House Price Prediction

1. Objective

Implement a linear regression model to predict house sale prices using:

- GrLivArea (above-grade living area)
- BedroomAbvGr (bedrooms above grade)
- TotalBath = FullBath + 0.5×HalfBath

2. Dataset

Kaggle “House Prices – Advanced Regression Techniques”

- Train: 1,460 rows, 81 columns
- No missing values in selected features

3. Methodology

a) Load train.csv → rename to dataset.csv

b) Engineer TotalBath

c) Train-test split (80/20, random_state=42)

d) Fit sklearn.LinearRegression

e) Evaluate with R^2, RMSE, MAE

4. Results (example run)

R^2 : 0.6398

RMSE : $70,124

MAE : $49,832

Coefficients:

GrLivArea ≈ $109.6 per sq ft

BedroomAbvGr≈ $15,200 per bedroom

TotalBath ≈ $39,800 per bath

Intercept ≈ -$48,300

5. Visualizations

- Actual vs Predicted scatter (model_plots.png)
- Residual plot (random scatter around zero)
- Correlation heatmap

6. Interpretation

- Square footage is the dominant driver (correlation 0.71 with SalePrice).
- Adding bedrooms & bathrooms improves R^2 from ~0.50 to 0.64.
- Model is simple, interpretable, and meets the task requirement.

7. Possible Extensions

- Log-transform SalePrice → optimize RMSLE (Kaggle metric)
- Include basement baths, garage area, year built

- Try Ridge/Lasso for regularization

8. Disclaimer
"This project is based on an open-source example from GitHub, modified to include additional features and improvements."

Author: Tsehaye Araya Hailemariam
Date: October 2025