Nama/NIM: Rafli Limandijaya/1103210243

1. Feature Engineering, EDA, dan Data Visualization

-EDA (Exploratory Data Analysis):

* Visualisasi distribusi semua fitur numerik menggunakan histogram dan KDE plot.
* Visualisasi korelasi antar fitur numerik dengan heatmap.
* Visualisasi fitur kategorikal menggunakan countplot (baik yang punya sedikit maupun banyak kategori).

-Feature Engineering:

* Penanganan missing values:
* Fitur numerik diisi dengan median.
* Fitur kategorikal diisi dengan modus (nilai terbanya

- Encoding kategorikal: Label encoding menggunakan .astype('category').cat.codes.

-Penambahan fitur baru:

* Fitur TotalArea = TotalBsmtSF + 1stFlrSF + 2ndFlrSF.

-Transformasi log untuk fitur yang skewed:

* Dilakukan log1p (np.log1p) pada fitur dengan skewness > 0.75 untuk mengurangi efek outlier.

2. **Matriks Evaluasi**

Model dievaluasi dengan tiga metrik utama:

* **RMSE (Root Mean Squared Error)**:

A black and white math equation

AI-generated content may be incorrect.

Mengukur rata-rata kesalahan prediksi. Nilai lebih kecil = prediksi lebih akurat.

* **MSE (Mean Squared Error)**:

A number and equation with numbers and symbols

AI-generated content may be incorrect.

Dipakai secara tidak langsung karena cross\_val\_score(..., scoring='neg\_mean\_squared\_error')

* **R² Score (Koefisien Determinasi)**:

A black symbols with white text

AI-generated content may be incorrect.

Menunjukkan seberapa baik model menjelaskan variasi data. Nilai 1 = sempurna.

3. Perbandingan model

* Saya menggunakan model linier regression dan random forest regression. Pengecekan overfitting dilakukan dengan membandingkan train RMSE dan validation RMSE. Didapatkan bahwa rasio Train/Val < 1 → artinya tidak overfit parah. Random Forest cenderung lebih akurat, tapi overfitting sedikit karena RMSE training-nya terlalu rendah.

- Training RMSE: 30504.67

-R²: 0.8525

-Baseline RMSE (mean): 79415.29  
Model saya jauh lebih baik daripada baseline (mean prediksi), artinya model ini sudah menangkap pola yang bermanfaat.

-Linear Regression:

* CV RMSE: 34968
* Train RMSE: 29065
* Val RMSE: 37446
* Overfitting Ratio: 0.776  
  Cukup bagus, tapi masih bisa ditingkatkan.

-Random Forest:

* CV RMSE: 30084
* Train RMSE: 11019 (sangat kecil)
* Val RMSE: 28379
* Overfitting Ratio: 0.388  
  Model Random Forest punya performa validasi yang jauh lebih baik, tapi sedikit overfitting karena train RMSE-nya terlalu rendah dibanding validasi.

R² Score Random Forest: 0.8950 → Lebih tinggi dari Linear Regression, artinya lebih akurat secara umum.

4.