📌 Project Documentation: Model Training & Evaluation (File 3)

1. Preparing Data for Model Training

After performing **feature selection** and **merging selected features**, the final dataset was saved as:
📄 **merged_to_keep_features.csv**

Data Cleaning Before Training

- **Dropped unnecessary columns**:
  - 'Unnamed: 0' (index column from CSV export)
  - 'Id' (unique identifier, not useful for modeling)
- Ensured the dataset contained only relevant **features for training**.

---

2. Implementing Multiple Linear Regression

To build the predictive model:

- Used **Multiple Linear Regression** as the modeling approach.
- The dataset was **split into training and test sets**:
  - **75% of the data** was used for training the model.
  - **25% of the data** was used for testing.
- The model was trained using the selected features to predict house prices.

---

3. Predicting and Evaluating Model Performance

After training, predictions were made on the **test set**, and the model's performance was evaluated.
A key part of this analysis was understanding the **importance of each feature** in the model.

- The **coefficients of the model** were extracted to rank features by importance.
- This helped in understanding which variables had the most impact on the target variable.

---

4. Model Evaluation Metrics

To measure how well the model performed, two key metrics were used:

Mean Squared Error (MSE)

📌 What is Mean Squared Error (MSE)?

- MSE calculates the **average squared difference** between **actual values** and **predicted values**.

- A **lower MSE** means the model's predictions are **closer** to the actual values.
- The obtained MSE:
  Mean Squared Error: 0.029322627971038136
  → **Indicates a relatively low error rate**, meaning the model makes accurate predictions with minimal error.

R-Squared Score (R^2)

📌 **What is R-Squared (R^2)?**

- R^2 measures **how well the independent variables explain the variability of the dependent variable**.
- Values range from **0 to 1**:
  - **1.0** → Perfect prediction.
  - **0.0** → Model does not explain any variance.
- The obtained R^2 score:
  R-squared: 0.8306588125579762
  → **Indicates that ~83% of the variance in the target variable is explained by the model**, meaning it captures most of the patterns in the dataset.

---

5. Saving the Model for Future Use

To allow future use without retraining, the **trained model was saved** using the **joblib** library in .pkl format.
This enables quick reloading and deployment of the model for predictions.

---

6. Summary of Work Done in File 3

✓ Merged selected features & saved them as merged_to_keep_features.csv
✓ Dropped unnecessary columns (Unnamed: 0, Id)
✓ Implemented Multiple Linear Regression
✓ Split dataset into 75% training, 25% testing
✓ Trained the model & made predictions
✓ Analyzed feature importance using model coefficients
✓ Evaluated model performance using MSE & R^2 Score
✓ Saved the trained model in .pkl format for future use