Dataset: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

1. Unzip the dataset

2. Load train.csv

3. Print:

- shape
- columns
- first 5 rows

4. Check missing values count per column

TASK 1 -> Exploratory Data Analysis (EDA)

Goal: Understand the structure, target distribution, and missing data.

1. Target distribution:
    - Plot distribution of SalePrice
    - Check outliers
2. Correlation with target:
    - Correlation heatmap of numeric ones
    - Top features correlated with SalePrice
3. Missing data analysis:
    - Percent missing per column
    - Decide if drop / impute

TASK 2 -> Build Preprocessing Pipeline

Goal: Create a pipeline to process numerical & categorical features.

The pipeline must:

- Handle numerical features
- Handle categorical features
- Impute missing values
- Scale or encode features

Split processing into blocks

1. Numeric pipeline
2. Categorical pipeline
3. Combine using ColumnTransformer

TASK 3 -> Train-Test Split

Goal: Create reliable evaluation.

1. Split dataset into train/test (80/20)
2. Use random_state=42

- X_train, X_test, y_train, y_test

TASK 4 -> Train Linear Regression Model

1. Create a full pipeline with preprocessing + LinearRegression
2. Train on X_train, y_train
3. Predict on X_test

TASK 5 -> Model Evaluation (MAE, RMSE, R2)