🏡 Data Science Project: House Price Prediction (Regression)

Skills: NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn
Goal: Build a machine learning model to predict house prices.

📘 📂 Dataset Columns (you will receive the CSV when you request it)

- **area** (square feet)
- **bedrooms**
- **bathrooms**
- **age_of_house** (years)
- **location_score** (0–10 rating)
- **price** (target)

📌 Part 1 — Data Exploration (EDA)

1. Load the dataset

- Load CSV using Pandas
- Show first 10 rows
- Print shape (rows, columns)

- Show dataset statistics (`df.describe()`)

---

2. Data Cleaning

- Check for missing values
- Check data types
- Identify outliers (boxplots recommended)
- Fix or remove incorrect values if found

---

3. Visual Analysis

Create the following visualizations using Matplotlib/Seaborn:

- Histogram of house price
- Distribution plot for area
- Scatter plot: area vs price
- Scatter plot: bedrooms vs price
- Correlation heatmap of all features
- Bar chart of average price by number of bedrooms

📌 Part 2 — Feature Engineering

4. Create new features

Examples you can add:

- `price_per_sqft = price / area`
- Categorize house age: new / medium / old
- Normalize or scale features if needed

📌 Part 3 — Machine Learning Model

5. Prepare the data

- Define **X** (all input features)
- Define **y** (price)

Split dataset into **train/test (80/20)**

```
 from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

-

---

6. Train the Regression Models

Students must train **at least one**, but ideally try **two**:

- Linear Regression
- Random Forest Regressor
- Decision Tree Regressor
- Ridge or Lasso regression

---

7. Evaluate the Models

Using Scikit-Learn:

- **MAE** (Mean Absolute Error)
- **MSE** (Mean Squared Error)
- **RMSE**
- **R^2 score**

Example:

from sklearn.metrics import mean_squared_error, r2_score

Plot:

- Actual vs Predicted prices (scatter plot)
- Residual plot

📌 Part 4 — Insights & Report

8. Write a short summary (8–12 sentences)

Include:

- Which features have strongest correlation with price
- Which model performed best and why
- What patterns you discovered in EDA
- Problems/limitations in the dataset
- Ideas to improve the model

📌 Submission Requirements

Students must submit:

1. **Jupyter Notebook (.ipynb)** with all code
2. All plots inside the notebook
3. A clean, readable workflow with comments
4. A short written summary