House Pricing Prediction

Supervisor:Abduallah Wagih

Members:

1-Ahmed Gamal -- 4230222
2-Abdelrheem Elsayed --4230754
3-Ahmed Mahmoud --4230800
4-Ibrahim Ashraf -- 4230920

This dataset contains data on residential properties, covering a range of features commonly used in house price prediction models.
As shown, it includes 2,000 rows and 9 columns (excluding the target variable):

Area (continuous): The size of the property in square feet

Bedrooms (integer): Number of bedrooms.

Bathrooms (integer): Number of bathrooms.

Floors (integer): Number of floors.

YearBuilt (integer): Year the house was built.

Condition (categorical): The current condition of the house (e.g., good, fair).

Garage (binary/integer): 1 if garage is available, otherwise 0.

Location (categorical): Different regions or neighborhoods.

Price (target, continuous): The sale price of the house.

Total size: 2,000 observations, ideal for applying regression and classification techniques.

Data Types & Missing Values:

Column	Data Type	Missing Values?
Area	float/int	None after cleaning
Bedrooms	int	None
Bathrooms	int	None
Floors	int	None
YearBuilt	int	None
Condition	object	None (after encoding)
Garage	object/int	None (after cleaning)
Location	object	None (after encoding)
Price	float	None

3. Statistical Summary
A quick numerical summary:

df.describe()

Area: Range from ~500 to ~5,000 sq ft (mean ≈ 2,000).

Bedrooms/Bathrooms: Commonly between 1 and 5.

YearBuilt: Mostly between 1900 and present.

Correlation heatmap reveals:

Positive correlation between Area & Price.

Modern houses usually fetch higher prices.

4. Data Preprocessing Steps:
1-Missing values: Handled with .fillna(0).

2-Encoding:

Condition, Garage, Location → numeric via LabelEncoder.

3-Scaling:

Used StandardScaler on continuous features (Area, Bedrooms, Bathrooms, Floors, YearBuilt, Garage) to normalize ranges.

5. Use Cases
Regression modeling — predicting house prices.

Feature analysis — understanding which factors influence prices most.

Categorical mapping — comparing prices across neighborhoods or home conditions.

6. Limitations & Future Enhancements:
Data size: 2,000 rows provides borderline performance for complex models.

Feature richness: Limited to basic numeric and categorical data—could be improved with features like "Proximity to schools" or "Renovation status".
Future improvements:
More rows or richer dataset
Feature engineering (e.g., house age, room density metrics)Advanced modeling (e.g., XGBoost, neural nets)
Hyperparameter tuning with tools like Grid.