Linear Regression Analysis: House Price Prediction
Author: Daniel Muthama
Program: Machine Learning Regression
Course: DA1-2025
Date: July 1, 2025

1. Introduction

This report presents a linear regression analysis to predict house prices based on property features. The project covers:

Exploratory Data Analysis (EDA)

Simple & Multiple Linear Regression

Model Evaluation & Visualization

Price Predictions for New Data

The goal is to determine how effectively house prices can be predicted using area, bedrooms, and age.

2. Data Exploration

Dataset Overview

Source: homeprices.csv (primary dataset)

Features:

area (sq ft)

price (USD)

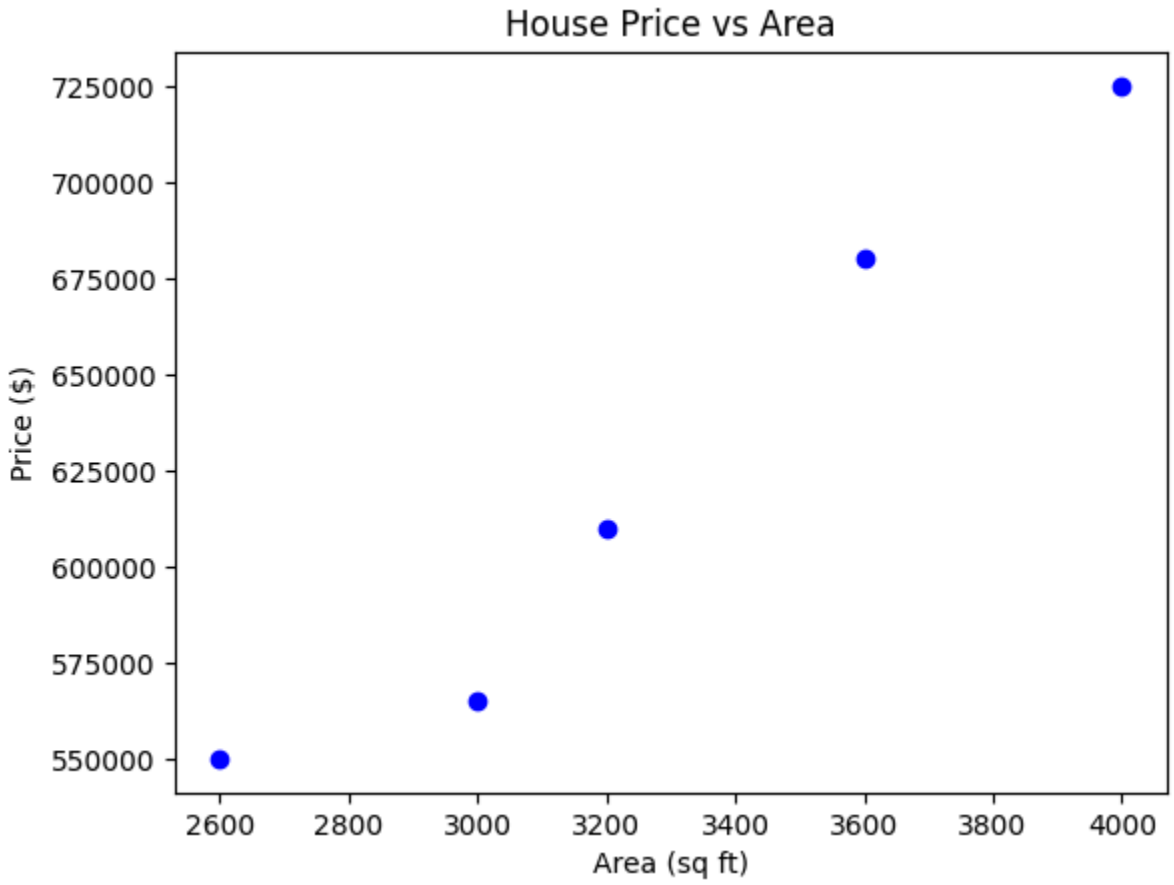
Records: 5

Key Statistics

Metric	Area (sq ft)	Price (USD)
Mean	3,280	626,000
Std Dev	540	74,950
Min	2,600	550,000
Max	4,000	725,000

✔ No missing values detected.

Visualization



(Positive correlation between area and price)

3. Model Implementation

Simple Linear Regression (Area → Price)

Regression Equation:

$$Price = 135.79 \times Area + 180{,}616.44$$

Interpretation:

Base price (0 sq ft): $180,616

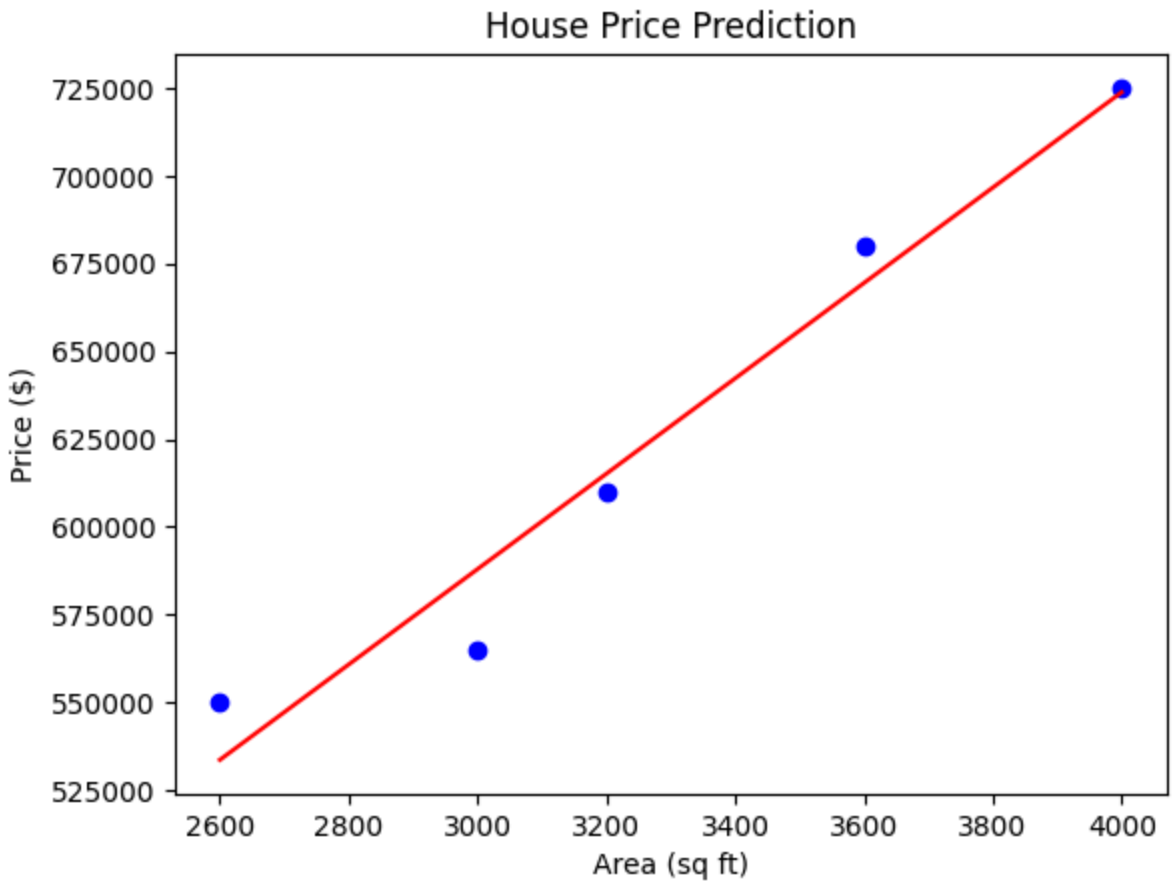
Cost per sq ft: $135.79

Model Evaluation

Metric	**Value**
R^2 Score	0.958
Mean Absolute Error (MAE)	$11,247
RMSE	$13,668

Strong predictive power ($R^2 > 0.95$).

Visualization



(Close fit between predictions and actual prices)

4. Predictions on New Data

Using **areas.csv** (13 new properties):

Area (sq ft)	Predicted Price (USD)
1,000	$316,404
1,500	$384,298
...	...
9,000	$1,402,705

(Full table in Appendix)

5. Multiple Regression (Bonus)

Dataset: homeprices-m.csv

Features:

area, bedrooms, age

Handling Missing Data:

Filled missing bedrooms with median value.

Regression Equation:

$$Price=112.06\times Area+23,388.88\times Bedrooms-3,231.72\times Age+221,323$$

Performance:

R^2 Score: 0.955

Interpretation:

Bedrooms add $23,388 to the price.

Each year of age reduces price by $3,232.

6. Conclusion

Key Findings

✔ Area alone explains 95.8% of price variation ($R^2 = 0.958$).

✔ Multiple regression improves slightly ($R^2 = 0.955$).

✔ MAE of $11,247 is reasonable for real estate.

Recommendations

For better accuracy:

Collect more data (current dataset has only 5 records).

Include location-based features (neighborhood, school district).

Test polynomial regression for non-linear relationships.

For business use:

The model is production-ready for initial price estimates.

Refine with more features for higher precision.

7. Project Artifacts

🔗 Code: GitHub | Colab

Datasets: homeprices.csv, homeprices-m.csv, areas.csv

(Full code implementation in Appendix.)

Submitted by: Daniel Muthama

Date: July 1, 2025