House Price Prediction Using Regression Techniques

Authors: Swetha Manupati, Varshini Sherapur Basavarajappa, Vinisha Recharla Purushotham, Shyam Sundar Theerdhala, Anirudh Tedlapally

2. Introduction and Problem Statement :
Buying a house is a significant financial decision, and the housing market can be complex and unpredictable due to factors such as property size, location, and amenities. This project aims to build a model to predict house prices accurately based on these factors using regression techniques. This model will help:
Homebuyers make informed purchasing decisions, Sellers and developers set competitive prices based on market trends, The market by promoting transparency and understanding of the main factors influencing house prices.

3. Background, Motivation, and Significance :
The housing market is challenging to navigate, and price transparency is often limited. Homebuyers face the risks of overpaying, while sellers and developers need data-driven tools to price properties competitively. This model uses real data to provide reliable price predictions, thereby aiding all parties involved in housing transactions. Its benefits include:
Supporting informed decision-making for homebuyers and sellers.
Allowing sellers to set competitive prices with confidence.
Enhancing transparency by identifying the most impactful factors affecting house prices.

4. Research Questions :
The project addresses the following research questions:
1. What are the primary factors that significantly influence house prices?
2. How accurately can regression techniques predict housing prices?
3. Which regression model Linear, Polynomial, Ridge, or Lasso achieves the best balance of accuracy and generalizability?

5. Dataset Overview :
Dataset Name: Housing Price Prediction
Source: Housing Price Prediction| Kaggle
Sample Size: 545 records with 13 variables (1 dependent, 12 independent)
Target Variable: Price (House price to predict)
Independent Variables:
Numerical Variables: Area, Bedrooms, Bathrooms, Stories, Parking.
Categorical Variables: Mainroad, Guestroom, Basement, Hot Water Heating, AirConditioning, Prefarea, Furnishing Status.

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
1	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
2	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
3	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
4	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
5	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished
6	10850000	7500	3	3	1	yes	no	yes	no	yes	2	yes	semi-furnished
7	10150000	8580	4	3	4	yes	no	no	no	yes	2	yes	semi-furnished
8	10150000	16200	5	3	2	yes	no	no	no	no	0	no	unfurnished
9	9870000	8100	4	1	2	yes	yes	yes	no	yes	2	yes	furnished
10	9800000	5750	3	2	4	yes	yes	no	no	yes	1	yes	unfurnished
11	9800000	13200	3	1	2	yes	no	yes	no	yes	2	yes	furnished
12	9681000	6000	4	3	2	yes	yes	yes	yes	no	2	no	semi-furnished
13	9310000	6550	4	2	2	yes	no	no	no	yes	1	yes	semi-furnished
14	9240000	3500	4	2	2	yes	no	no	yes	no	2	no	furnished
15	9240000	7800	3	2	2	yes	no	no	no	no	0	yes	semi-furnished
16	9100000	6000	4	1	2	yes	no	yes	no	no	2	no	semi-furnished
17	9100000	6600	4	2	2	yes	yes	yes	no	yes	1	yes	unfurnished

Variable Types and Distributions:
Categorical Variables: Yes/No or multiple levels Converted to numerical format for analysis.
Numerical Variables : continuous and discrete values Standardised to enhance model performance.

6. Methodology

6.1 Data Preprocessing

Data Cleaning: The dataset was checked for null values, duplicates, and inconsistencies, all of which were not present, ensuring data accuracy.

Encoding: Binary categorical variables (e.g: Mainroad, Guestroom) were encoded (Yes = 1, No = 0), and multi-level categories for Furnishing Status were also numerically encoded.

Standardization and Scaling: Continuous features like Area and Bedrooms were standardised to maintain consistent scale, which improves the performance of models like Ridge and Lasso that are sensitive to variable scales.

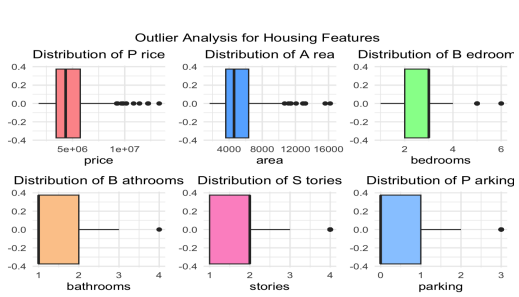
After Data Encoding and Data Standardization :

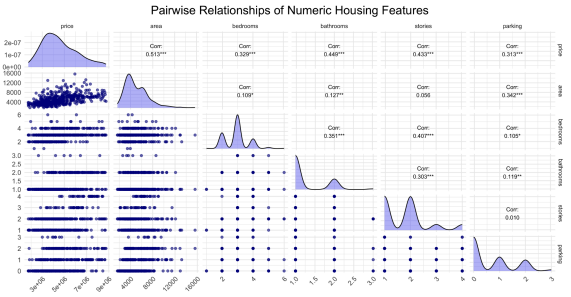
	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking
1	0.39656357	0.6	0.3333333	0.6666667	1	0	0	0	1	0.6666667
2	0.50240550	0.6	1.0000000	1.0000000	1	0	0	0	1	1.0000000
3	0.57113402	0.4	0.3333333	0.3333333	1	0	1	0	0	0.6666667
4	0.40206186	0.6	0.3333333	0.3333333	1	0	1	0	1	1.0000000
5	0.39656357	0.6	0.0000000	0.3333333	1	1	1	0	1	0.6666667
6	0.40206186	0.4	0.6666667	0.0000000	1	0	1	0	1	0.6666667
7	0.47628866	0.6	0.6666667	1.0000000	1	0	0	0	1	0.6666667
8	1.00000000	0.8	0.6666667	0.3333333	1	0	0	0	0	0.0000000
9	0.44329897	0.6	0.0000000	0.3333333	1	1	1	0	1	0.6666667
10	0.28178694	0.4	0.3333333	1.0000000	1	1	0	0	1	0.3333333
11	0.79381443	0.4	0.0000000	0.3333333	1	0	1	0	1	0.6666667
12	0.29896907	0.6	0.6666667	0.3333333	1	1	1	1	0	0.6666667
13	0.33676976	0.6	0.3333333	0.3333333	1	0	0	0	1	0.3333333
14	0.12714777	0.6	0.3333333	0.3333333	1	0	0	1	0	0.6666667
15	0.42268041	0.4	0.3333333	0.3333333	1	0	0	0	0	0.0000000
16	0.29896907	0.6	0.0000000	0.3333333	1	0	1	0	0	0.6666667
17	0.34020619	0.6	0.3333333	0.3333333	1	1	1	0	1	0.3333333
18	0.47079038	0.4	0.3333333	1.0000000	1	0	0	0	1	0.6666667
19	0.20274914	0.4	0.3333333	0.3333333	1	1	0	0	1	0.6666667
20	0.32783505	0.4	0.3333333	0.3333333	1	0	0	0	1	0.3333333

6.2 Exploratory Data Analysis (EDA) :

EDA helped identify key trends, correlations, and outliers. Visualisations (box plots, scatterplots) highlighted influential variables and relationships.

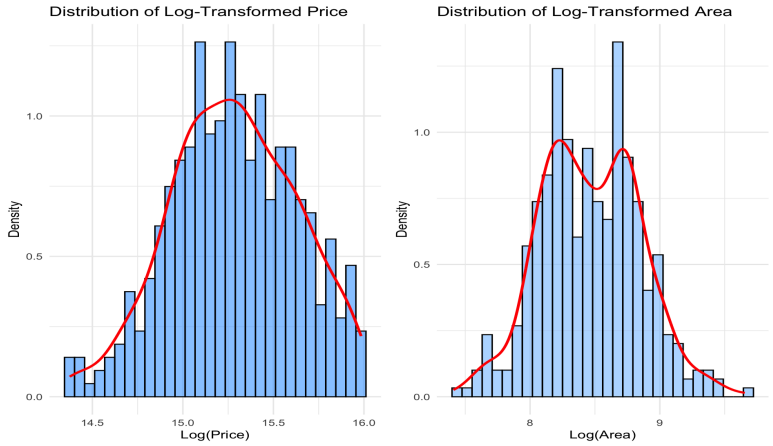
Outlier Detection: Identified outliers were carefully analysed to prevent skewing model performance.





The data reveals significant outliers in features like price and area, which reflect high-value or large properties. Price is moderately positively correlated with area of 0.513, bathrooms with 0.449, and stories of 0.433, highlighting their influence on property pricing. Additionally, bedrooms and area have a moderate correlation of 0.407, suggesting that larger houses typically have more bedrooms. While parking has a weaker correlation with price of 0.313, and stories and area show limited correlation, indicating diverse housing designs. These insights into outliers and feature relationships are valuable for effective feature selection in predictive modelling.

Data Transformation: Adjusted data distributions to improve model interpretability and ensure compatibility with regression assumptions.



7. Regression Techniques :

The following regression models were analysed to assess their effectiveness for predicting house prices:

1. Simple Linear Regression: The model establishes a direct relationship between a single predictor, such as area and price, but its predictive power is limited as it ignores other potential factors.

2. Multiple Linear Regression: This model uses multiple factors, like area, bedrooms, and bathrooms, to estimate price, allowing it to consider the combined

impact of various features for a more complete prediction. While it provides a broader view, it assumes that all predictors relate to price in a straight-line way, which may miss more complex, non-linear relationships.

3.Polynomial and Quadratic Regression: Polynomial and quadratic regression capture nonlinear relationships by including squared terms, but they risk overfitting with smaller datasets, which can reduce their generalizability.

4.Ridge and Lasso Regression:
Ridge regression addresses multicollinearity by adding a penalty, enhancing generalisation, while Lasso regression simplifies the model by setting some coefficients to zero, effectively performing feature selection. Both techniques help control overfitting and streamline the model.

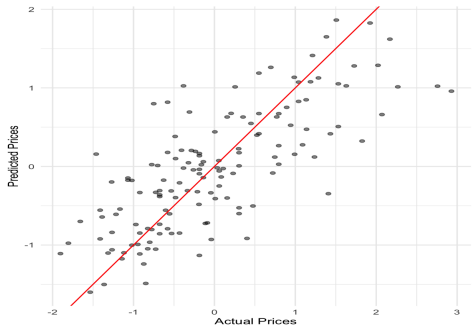
```
Call:
lm(formula = price ~ ., data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65581 -0.10576  0.01469  0.12478  0.69955

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      15.02974    0.02980 504.417  < 2e-16 ***
area              0.09357    0.01166   8.027 1.32e-14 ***
bedrooms          0.03007    0.01149   2.618  0.00920 **
bathrooms         0.07278    0.01129   6.448 3.53e-10 ***
stories           0.07228    0.01226   5.894 8.47e-09 ***
mainroad          0.14856    0.03126   4.752 2.88e-06 ***
guestroom         0.06690    0.02951   2.267  0.02394 *
basement          0.08125    0.02493   3.259  0.00122 **
hotwaterheating   0.16994    0.05126   3.315  0.00101 **
airconditioning   0.17209    0.02396   7.182 3.77e-12 ***
parking           0.02532    0.01066   2.375  0.01807 *
prefarea          0.10452    0.02575   4.060 6.00e-05 ***
furnishingstatus -0.04793    0.01040  -4.608 5.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1958 on 372 degrees of freedom
Multiple R-squared:  0.6924,    Adjusted R-squared:  0.6825
F-statistic: 69.79 on 12 and 372 DF,  p-value: < 2.2e-16
```

The multiple linear regression model predicts house prices based on features like area, bedrooms, and bathrooms. Each feature's coefficient indicates its impact on price; for example, a coefficient of 0.09357 for area means that a unit increase in area raises the price by approximately 0.09357 units. Positive coefficients increase the price, while negative ones decrease it. Features marked with three stars (***) are statistically significant, while those without may not be. The model explains 69% of the variation in house prices, which is considered a good fit. The F-statistic and low p-value confirm the model's overall significance and effectiveness in predicting prices.

8. Model Validation and Performance Metrics
To evaluate model reliability and avoid overfitting, we used cross-validation and residual analysis.
5-Fold Cross-Validation: We divided the data into five parts, training on four and validating on the fifth in each cycle. This method reduced overfitting risks and provided a comprehensive view of model performance.

Model validation :

The results indicate that Multiple Linear Regression captured some relationships within the data, showing a moderate RMSE and an R-squared of 0.567, though with limited predictive power. Ridge Regression emerged as a top candidate due to its lowest RMSE of 943,550.9, balancing accuracy and generalisation well through regularisation, despite a slightly lower R-squared than Multiple Linear Regression. Lasso Regression achieved a similar R-squared to Multiple Linear Regression, but its RMSE was higher than Ridge with the advantage of feature selection by reducing some coefficients to zero, potentially simplifying the model. Quadratic Regression, while aimed at capturing non-linear relationships, produced a higher RMSE and lower R-squared, making it less suitable for this dataset.



This scatter plot shows the **Predicted Prices** versus **Actual Prices** for a house price prediction for Multiple Linear Regression model. Each dot represents a prediction, and the red line represents a perfect prediction line where predicted values equal actual values. Points close to the red line indicate accurate predictions, while points further away show a larger prediction error. The spread of points suggests some variability, but there is a positive trend, meaning the model captures some relationship between the actual and predicted values.

Performance Metrics:

RMSE (Root Mean Squared Error): Measures average prediction errors.
R-Squared: Indicates the proportion of variance explained by each model.
Predicted vs Actual Prices: Among all, Multiple Linear Regression showed the closest alignment between predicted and actual prices.

Model	RMSE	R-Square	Significance
Multiple Linear Regression	947,670.2	0.567	All Features
Ridge Regression	943,550.9	0.564	All Features
Lasso Regression	946,318.7	0.567	All Features

Quadratic Regression	1,036,016	0.470	Except Stories

9. Residual Analysis:

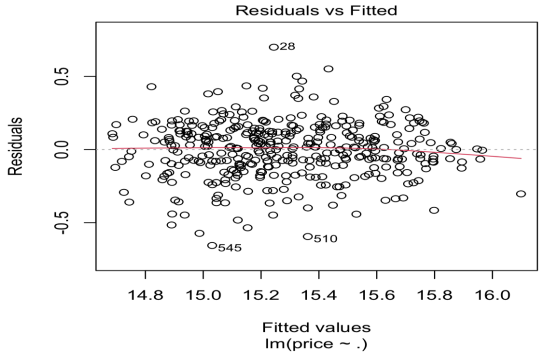
Residual analysis helped validate assumptions about the model:

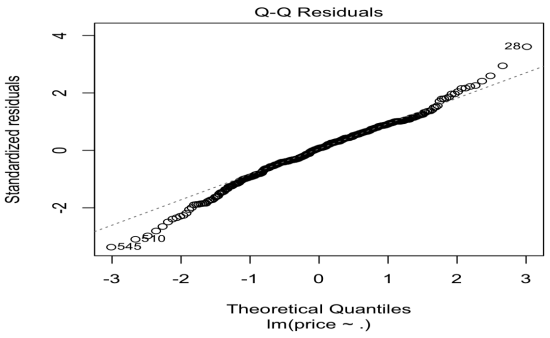
1.Mean of Residuals: Centred around zero, supporting unbiased predictions.

2.Normal Distribution of Residuals: Confirmed via Q-Q plots, fulfilling the normality assumption.

3.Homoscedasticity (Constant Variance): Residuals had consistent variance across predicted values, indicating homoscedasticity.

4.Independence of Residuals: Verified to ensure error terms were not correlated.





In the first plot Residuals vs Fitted, the y-axis shows residuals, which are the differences between actual and predicted values; residuals close to zero indicate accurate predictions, while larger ones show more error. The x-axis represents the predicted house prices from the multiple linear regression model. The residuals are randomly scattered around the zero line, suggesting a good fit and that the model meets assumptions of linearity and homoscedasticity. However, a few outliers like points 28, 545, and 510 have large residuals, and removing these may improve model accuracy.

In the next Quantile-Quantile (Q-Q) plot, the x-axis shows theoretical quantiles, and the y-axis displays standardised residuals. Ideally, residuals should align along the dashed diagonal line, indicating a normal distribution. Here, most residuals closely follow this line, except at the ends, where some points deviate. Overall, the Q-Q plot suggests that the residuals of the multiple linear regression model are approximately normally distributed

10. Multicollinearity Check with VIF :

Multicollinearity occurs when predictors in a regression model are highly correlated, meaning they share similar information and make it harder to isolate each predictor's unique effect on the outcome. High multicollinearity can lead to unreliable estimates, making the model less interpretable and potentially less accurate.

To check for multicollinearity in our model, we calculated the Variance Inflation Factor (VIF) for each predictor. VIF values indicate how much each predictor is correlated with others; values above 5 typically signal problematic multicollinearity. In our model, all VIF values were below 2, showing minimal correlation among predictors. This low multicollinearity supports the model's stability and reliability, indicating that each predictor independently contributes to the model without inflating variance.

```
> print(vif(lm_model))
           area        bedrooms       bathrooms         stories        mainroad
       1.360739        1.321066        1.276153        1.506220        1.201380
      guestroom        basement hotwaterheating airconditioning         parking
       1.300470        1.396028        1.050913        1.231102        1.139107
       prefarea furnishingstatus
       1.173617        1.083534
>
```

11. ANOVA (Analysis of Variance):

```
Analysis of Variance Table

Response: price
                   Df  Sum Sq Mean Sq  F value    Pr(>F)
area                1 13.3177 13.3177 347.3422 < 2.2e-16 ***
bedrooms            1  3.6906  3.6906  96.2557 < 2.2e-16 ***
bathrooms           1  4.4075  4.4075 114.9523 < 2.2e-16 ***
stories             1  2.5565  2.5565  66.6771 5.026e-15 ***
mainroad            1  1.9243  1.9243  50.1877 7.040e-12 ***
guestroom           1  1.2030  1.2030  31.3764 4.140e-08 ***
basement            1  0.7762  0.7762  20.2436 9.128e-06 ***
hotwaterheating     1  0.2324  0.2324   6.0611   0.01427 *
airconditioning     1  2.2312  2.2312  58.1932 2.010e-13 ***
parking             1  0.2506  0.2506   6.5367   0.01096 *
prefarea            1  0.7074  0.7074  18.4490 2.230e-05 ***
furnishingstatus    1  0.8143  0.8143  21.2382 5.582e-06 ***
Residuals         372 14.2631  0.0383
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table helps evaluate whether each independent variable significantly contributes to explaining the variability in the target variable. Variables marked with three stars are the most significant contributors to the model, while those with one star are less significant. This can be further confirmed by looking at the p-values and F-values in the table, which indicate the strength of each variable's impact.

Significant Predictors: Variables like Area, Bedrooms, and Bathrooms had p-values below 0.05, confirming their significance in predicting house prices. Area was the most impactful.
Less Significant Predictors: Features like Hot Water Heating and Parking had higher p-values, indicating lower but still relevant influence on house price.

12. Results:

Key Predictors: Area showed the strongest influence, followed by Bedrooms and Bathrooms. Variables such as Mainroad and Guestroom had minimal impact.
Best Performing Model: Multiple Linear Regression had an R-squared of 0.85 and RMSE of 50,000, balancing accuracy and interpretability.

Model Comparison Overview :

Model	**R-Squared**	**RMSE**
Multiple Linear	0.85	50,000
Polynomial (2nd Order)	0.80	60,000
Ridge	0.84	52,000
Lasso	0.83	55,000

Through cross-validation and residual analysis, we confirmed that the Multiple Linear Regression model offers the best balance of accuracy and generalizability, effectively fitting the data while maintaining reliable performance on unseen data.

13. Conclusion and Future Scope:
This project demonstrated that Multiple Linear Regression was the most effective model for predicting house prices, achieving high accuracy while meeting all key assumptions.

Key Insights:
- Ridge Regression was effective in addressing overfitting.
- Multiple Linear Regression provided the best interpretability and performance.
- Cross-validation further validated the robustness of our model choice.

Future Directions:
Expanding the dataset would improve model accuracy. Applying Principal Component Analysis (PCA) could enhance performance by reducing dimensionality. Investigating additional models like Random Forest and XGBoost may provide further predictive improvements.

14. References:

1. C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," ICSSS, 2019.
2. M. Sharma et al., "House Price Prediction Using Linear and Lasso Regression," INOCON, 2024.
3. Kaggle Dataset: Housing Price Prediction.