DATA ANALYSIS & REGRESSION – DSC 423

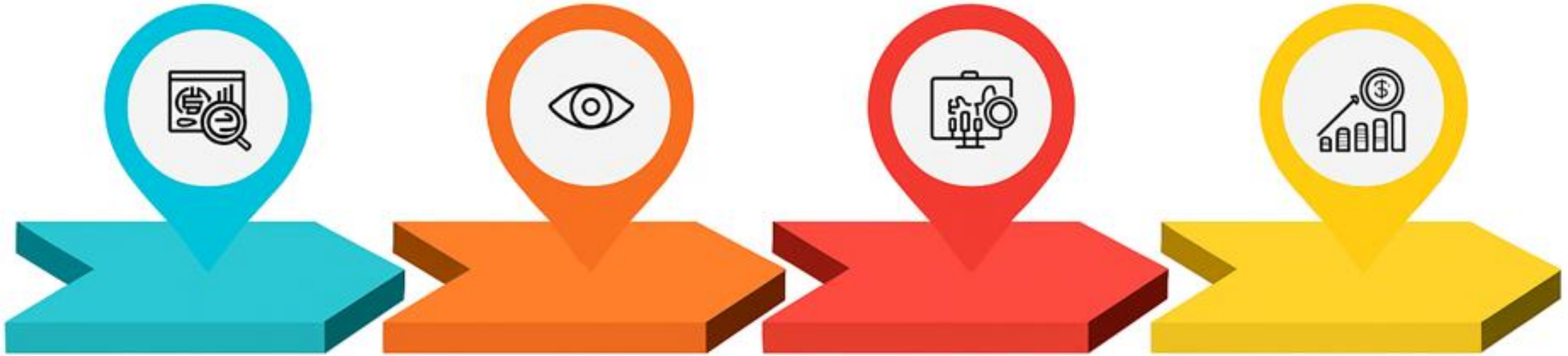
Divyarajsinh Chavda
2234740

Housing Market Data

Divyarajsinh Chavda

Student ID: 2234740

Market Research for Business Plan



Housing Market Research

The data explored and exxamin ed to understand dynamics within the housing market.

Interaction Term

A combined effect involve the Sales and Discount d that will be used for Model 2.

Model Evaluation

A thorough assessment and analysis of statistical models

Strategy Development

A strategic plan aimed at in increasing the sales and pro- fits in thhe housing sector

What factors can be analyses to better predict property price performance in the housing market?

In the context of the housing market, especially within Bengaluru's residential sector, predicting property prices involves analysing a set of structural and spatial features that influence home values. Core variables such as total square footage (total_sqft), number of bathrooms (bath), and number of bedrooms (bhk) provide quantitative insights into a property's size and livability, which are directly related to its price.

To capture both direct and interactive effects, we applied various regression models including Linear Regression, Interaction Terms, Polynomial Regression, Ridge, and Lasso Regression. By evaluating combinations such as total_sqft × bath, we could understand how the relationship between space and amenities contributes to the pricing trend.

Each model offered unique perspectives:

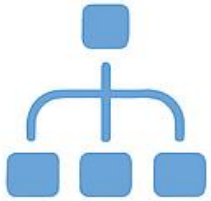
- Linear Regression revealed straightforward relationships.
- Interaction Models exposed how combined features behave together.
- Polynomial Regression captured non-linear patterns.
- Ridge and Lasso helped manage multicollinearity and optimised feature selection.

This analysis not only improves the accuracy of price prediction but also allows urban developers, buyers, and policy-makers to make data-informed decisions about real estate planning and investment.

About the Dataset

Urban Housing Insights

The dataset features residential housing data from Bengaluru, focusing on aspects like total square footage, nurbher of bedrooms (BHK), bathrooms, and property price



Smart Split Strategy

The data wa split using an 80 % training and 20 % testing approach to ensure accurate evaluation of model performance.

Analytical Blueprint

Multiple regression models were developed to predict housing prices, uncovering patterns in buyer preferences and pricing dynamics.

Diverse Model Arsenal

Techniques applied include Linear Regression, Polynomial Regression, Ridge. Lasso, and Interaction models to capture both linear and complex relationshi.

Real-World Accuracy Check

The best model achieved an R^2 of approximately 47%, showing the importance of data preprocessing and hinting at potential perfmance gains with location-based or catgonical variables

Exploratory Analysis

- Visual inspections were conducted using scatter plots, histograms, and box plots to observe the distribution of key variables such as Total-Sqft, bath, BHK, and price.

- Correlation analysis was applied to measure linear relationships between features. This revealed that
- Total- Sqft had the highest positive correlation with housing prices.

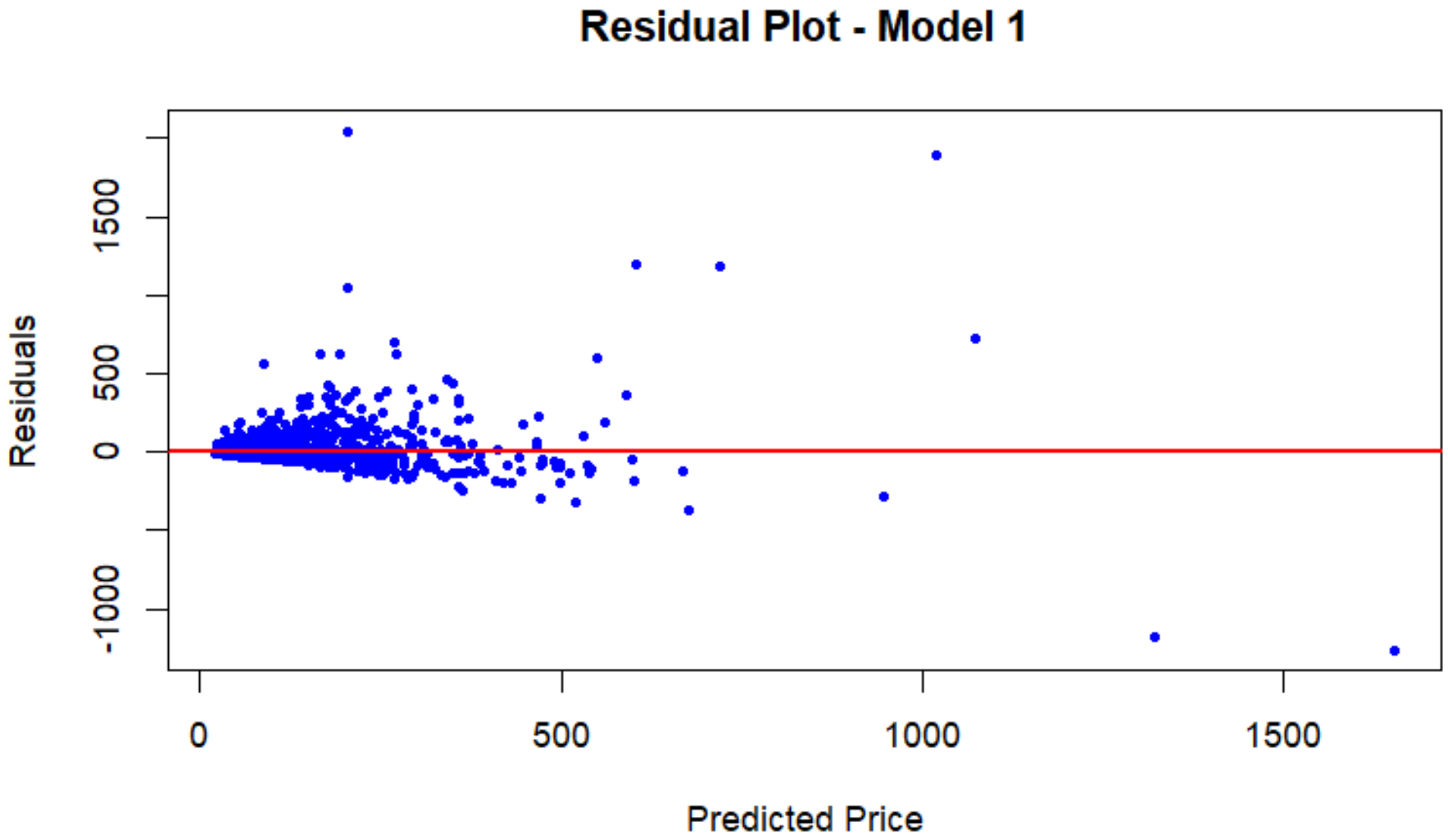
- Outlier detection was essential—extreme values in BHK (e.g., >10) or unrealistic square footage were removed to prevent distortion in model training.

- Feature transformation included extracting numeric values from categorical entries (e.g., converting "2 BHK" to 2), and converting ranges in total-Sqft into mean values.

- Missing values in important fields like price, bath, and size were handled through filtering to ensure clean and consistent input data.

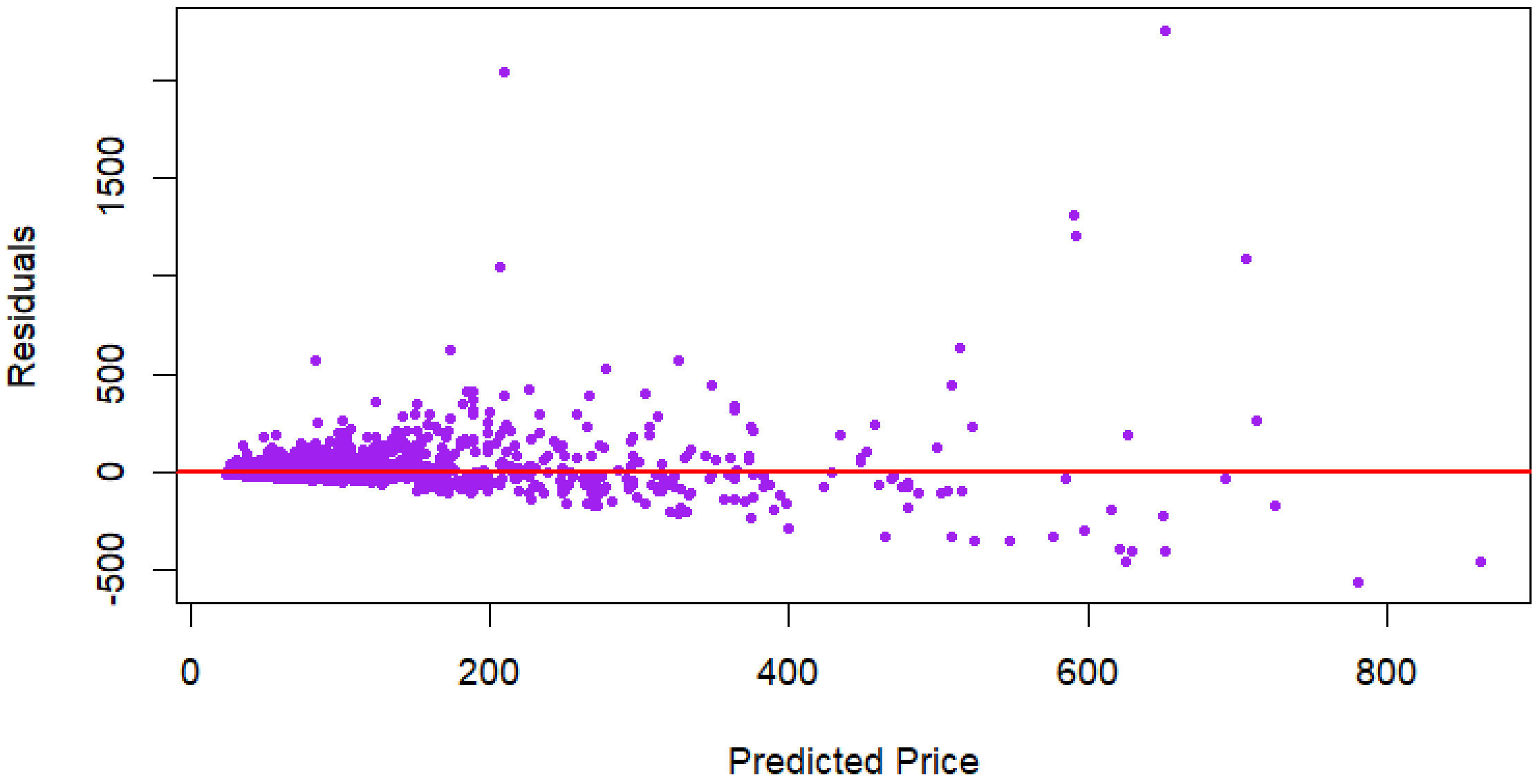
- Overall, this phase ensured that the dataset was properly structured, cleaned, and ready for meaningful regression analysis.

Linear Regression



The Linear Regression model was evaluated using the following metrics

R^2 (Test) : 47.28%
MAE: 37.18
Mean Squared Error (MSE): 10311
Root Mean Squared Error (RMSE): 101.54


Residual Plot - Model 3 (Polynomial)
Residuals
1500
500
0
-500
0
200
400
600
800
Predicted Price

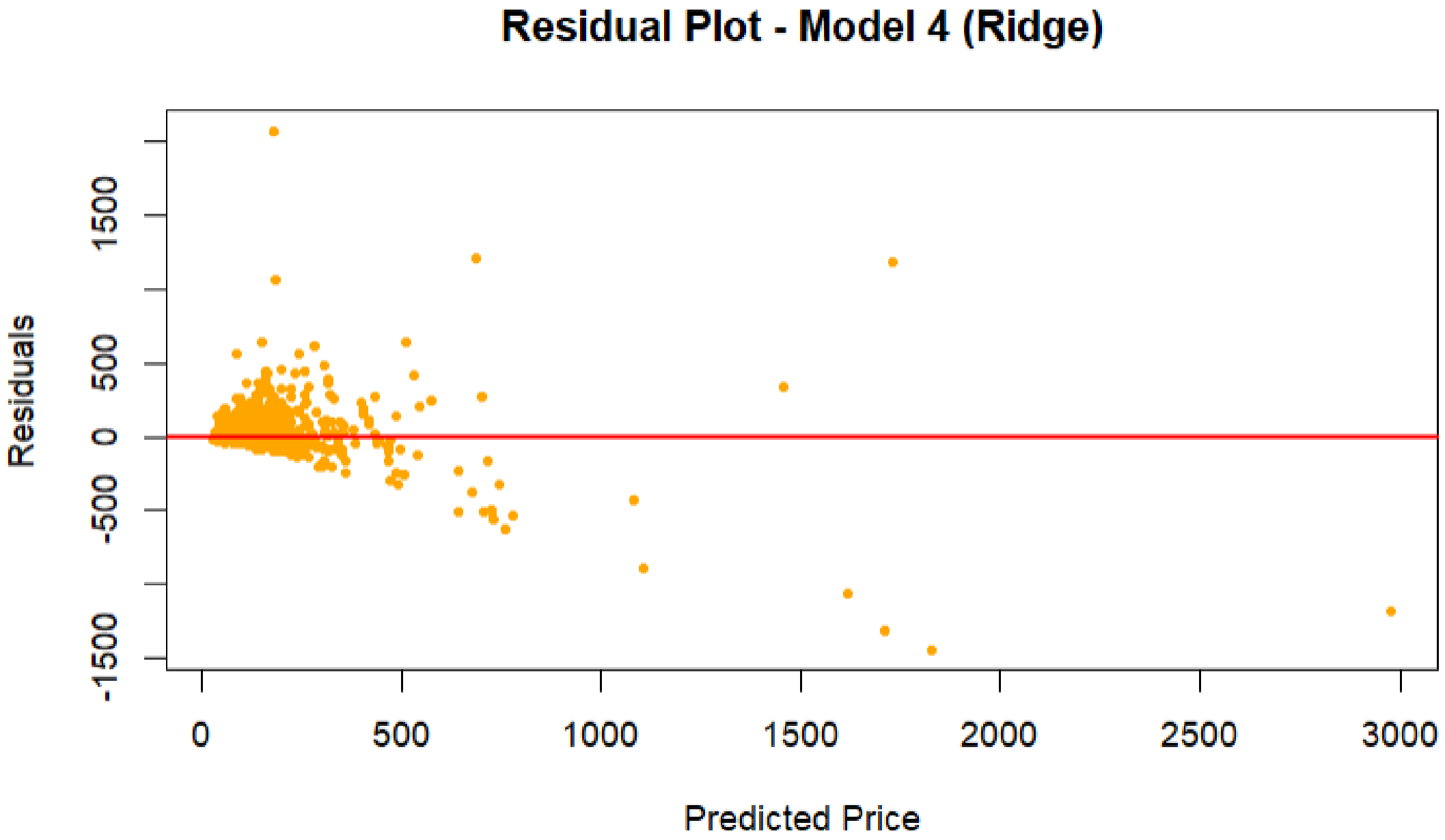
Polynomial Regression – Actual vs Predicted Price

Polynomial Regression incorporates non-linear features by including squared terms such as total_sqft², bath², and bhk² to model curvature in the relationship between house characteristics and price.

This model helps capture complex trends that cannot be explained by a simple straight-line relationship.
The residual plot shows the distribution of errors between actual and predicted prices, with most points lying close to the baseline, indicating reasonable performance.

Some variation is observed in higher predicted price ranges, but the model remains stable across most data points.
Polynomial terms allow the model to better fit data where price growth accelerates with increasing size or amenities.

Overall, this model balances flexibility and interpretability by introducing non-linearity in a controlled manner.
Model Evaluation Metrics:
R^2 Score: 46.49%
Mean Absolute Error (MAE): 37.80
Mean Squared Error (MSE): 10,465.47
Root Mean Squared Error (RMSE): 102.30


Residual Plot - Model 4 (Ridge)
Residuals
1500
500
0
-500
-1500
0
500
1000
1500
2000
2500
3000
Predicted Price

Ridge Regression – Actual vs Predicted Price

Ridge Regression uses **L2 regularisation**, which reduces overfitting by penalising large coefficients without eliminating features.

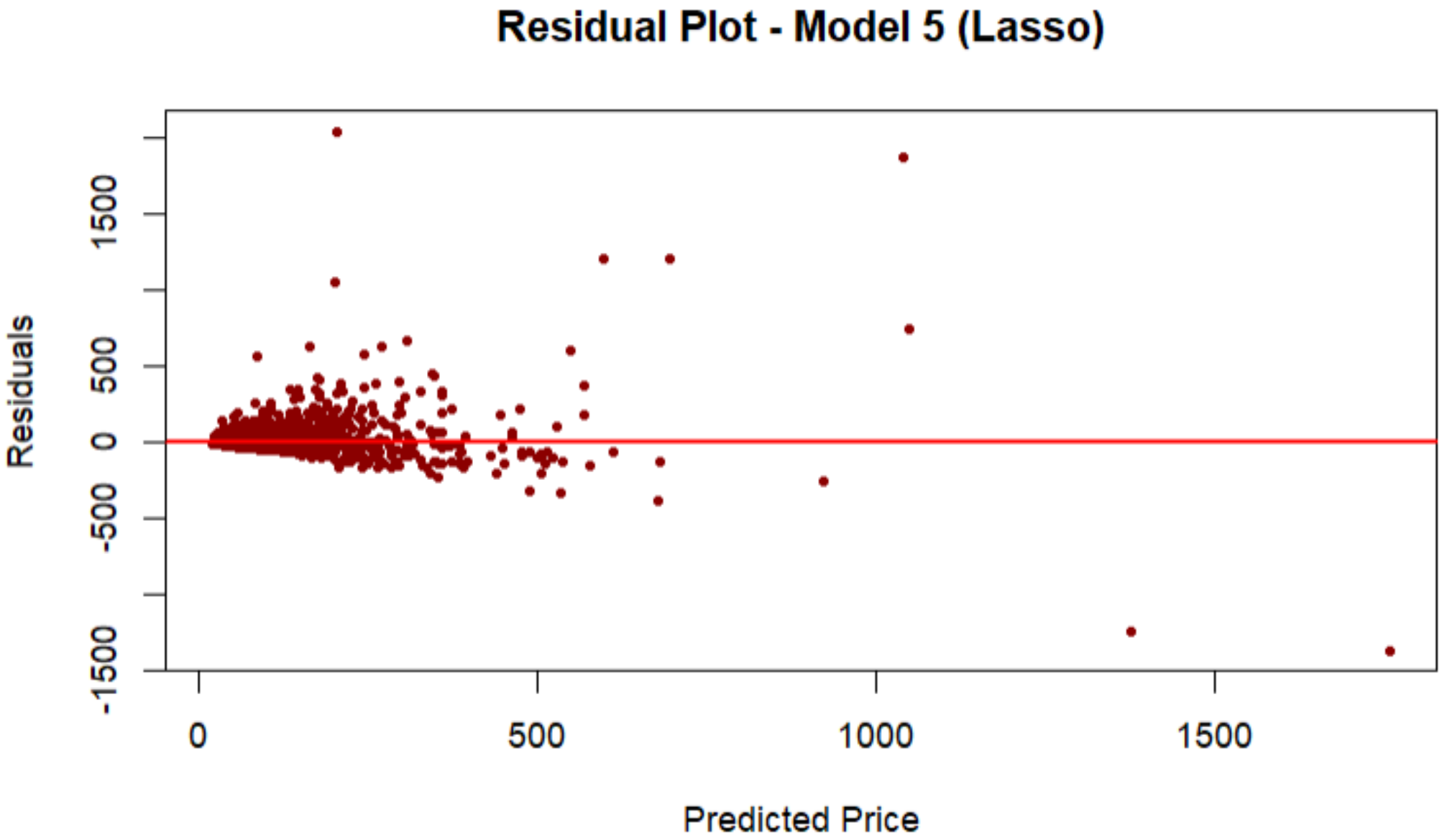
The **red line** on the residual plot shows the ideal prediction line where actual price equals predicted price.

Many predicted values align closely with the red line, but **some deviation is noticeable**, especially at higher price points.

The model **stabilised coefficient values** for features like total-sqft, BHK, and bath, preventing multicollinearity issues.

Tuning the regularisation strength (lambda) helped strike a balance between **model complexity and generalisation**.

Ridge provided a **more stable prediction range**, though it did not outperform simpler models in accuracy.


Residual Plot - Model 5 (Lasso)
Residuals
1500
500
0
-500
-1500
0
500
1000
1500
Predicted Price

Lasso Regression – Actual vs Predicted Price

Lasso Regression applies **L1 regularisation**, which helps reduce overfitting while also performing **automatic feature selection**.

The **red diagonal line** on the residual plot shows the ideal case where predicted price equals actual price.
Most observations lie near this line, suggesting **decent prediction accuracy** and moderate bias.
Lasso **shrinks the coefficients** of less important features (like weak variations in Bhk or bath) to zero, simplifying the model.

This leads to a **cleaner and more interpretable model**, particularly useful when dealing with limited numeric inputs.
It proved effective in narrowing down to the most impactful features, especially **total-sqft**, which consistently influenced price.

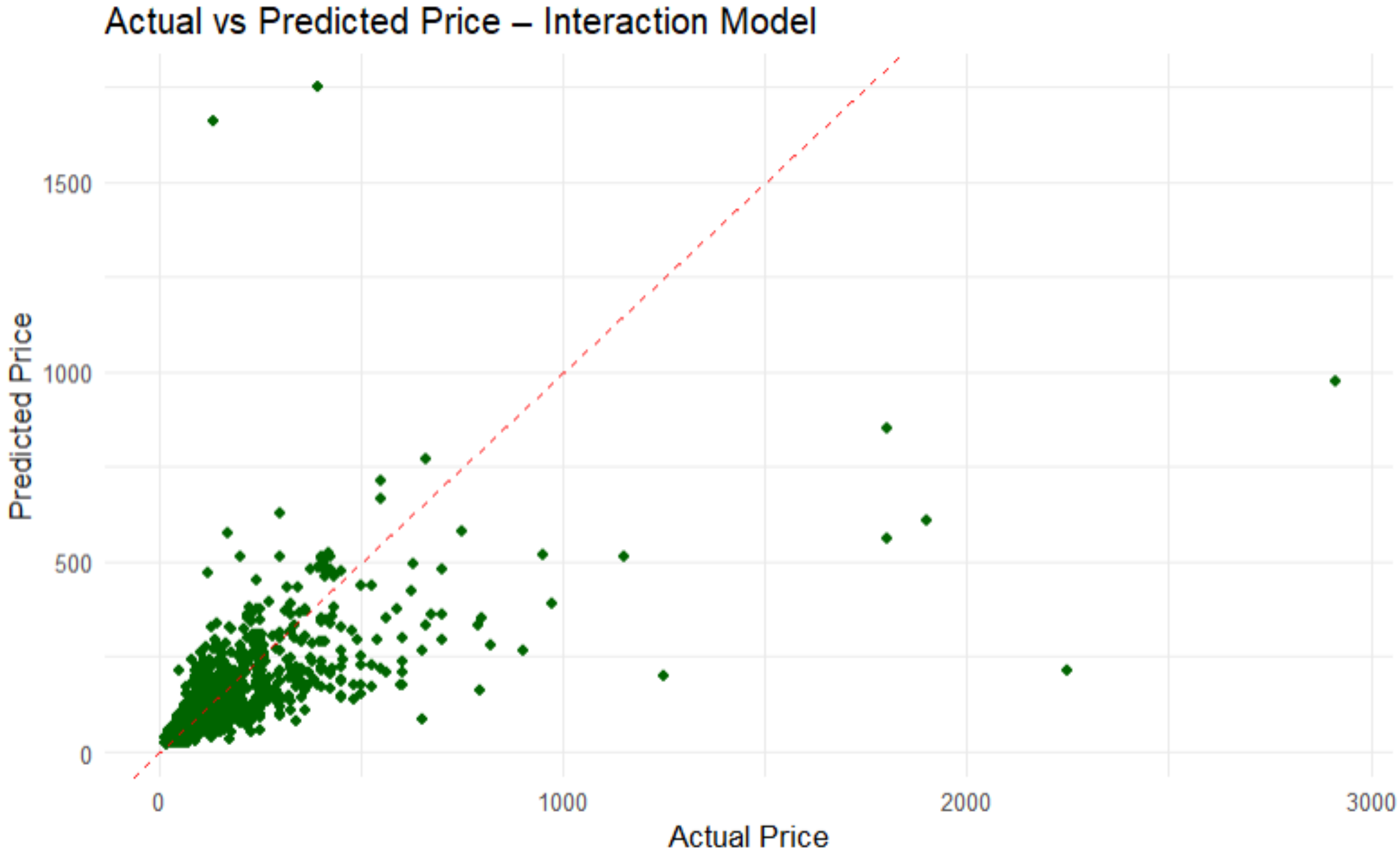
While the R^2 score was moderate, the Lasso model maintained a good balance between performance and simplicity.

Your Results Summary:
MSE: 10,416.70
RMSE: 102.06
MAE: 37.08
R^2 Score: 46.74%

Ridge vs. Lasso Regression: Performance Comparison

🔶 **Ridge Regression**
RMSE: 106.64
R-squared (Accuracy): 41.85%
MAE: 39.83
Ridge Regression used L2 regularisation to reduce the magnitude of coefficients and manage multicollinearity.
It retained all predictors but provided modest accuracy and slightly higher error values.
Suitable when you want to stabilise the model without completely removing any variable.

🔷 **Lasso Regression**
RMSE: 102.06
R-squared (Accuracy): 46.74%
MAE: 37.08
Lasso Regression applied L1 regularisation, which helped shrink less important features to zero.
This resulted in a more compact and interpretable model with slightly better accuracy than Ridge.
Ideal when feature selection is important, along with performance.


Actual vs Predicted Price – Interaction Model
Predicted Price
0
500
1000
1500
0
1000
2000
3000
Actual Price

Interaction Model – Actual vs Predicted Price

This model includes interaction terms such as **total_sqft × bhk** and **total_sqft × bath** to capture how features jointly affect house prices.

It helps identify how the layout and size combinations influence pricing in real estate.

The red diagonal line in the plot represents the **ideal prediction line**, where predicted price equals actual price.

Most points align near the line, but variation increases for high-priced homes.

This model enhances understanding of **conditional feature relationships**, especially when multiple property attributes scale together.

Useful in housing scenarios where **larger homes with more amenities** don't always increase price linearly.

Overcoming Housing Market Challenges

About Dataset

- Bengaluru housing slas data
- Variables include total_sq.ft, bhk, bath, price
- Interaction and no-linear models

Challenges

- High compestition in city zones
- Variable demand by location
- Price affected by new projects

Key Initiative

- Enhanced interaction model
- Polynomial regression applied
- Improved interpretability

Business Impact

- Captured compound effects
- Boosted model accuracy
- Generated actionable pricing insights

Final Thoughts

Regression models effectively helped identify key drivers of housing prices, such as total square footage, BHK, and number of bathrooms.

The Lasso Regression model provided the best balance between simplicity and accuracy, making it the most practical model for real-world applications.

The enhanced interaction model demonstrated that combining features (like total-sqft × BHK) can moderately improve

prediction performance by capturing the effects of layout and design structure.

Despite acceptable results, prediction accuracy remains limited due to the absence of categorical variables like location, property type, or society features, which play a significant role in real estate pricing.

Future Improvements

Incorporate location-based and categorical features to better reflect market influences on pricing.

Apply outlier detection and removal techniques more rigorously to minimise distortion in regression outcomes.

Test and compare non-linear models such as Random Forest or XG Boost to improve R^2 beyond the current limits.

Develop an interactive dashboard using tools like R Shiny or Power BI to visualise predictions and allow user-based input in real time.