Housing Case Study: Multiple Linear Regression

Problem Statement

A real estate company in Delhi wants to:

- Identify key factors affecting housing prices (e.g., area, number of bedrooms, parking).
- Build a linear regression model to predict house prices based on those factors.
- Evaluate model accuracy to determine how well the variables predict prices.

Tools and Libraries Used

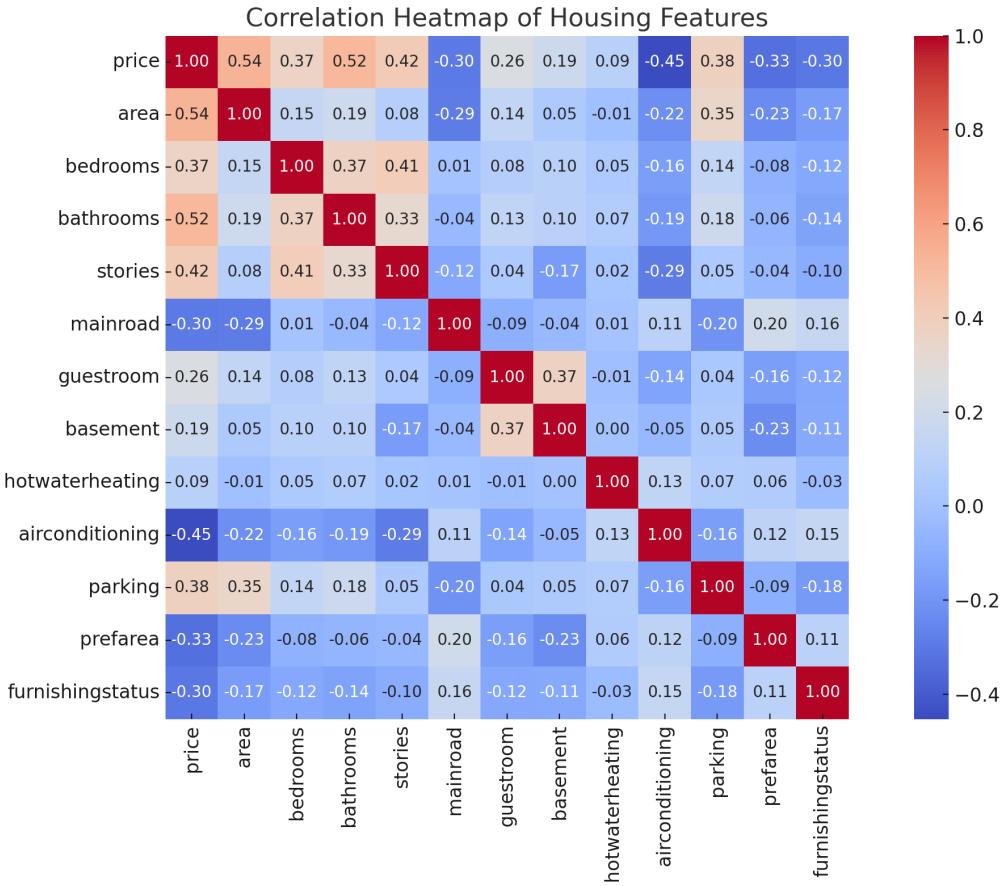
- NumPy: Numerical computations
- Pandas: Data handling and preprocessing
- Matplotlib & Seaborn: Data visualization
- Scikit-learn: Linear regression model and evaluation
- Statsmodels: Statistical analysis and diagnostics

Approach and Key Steps

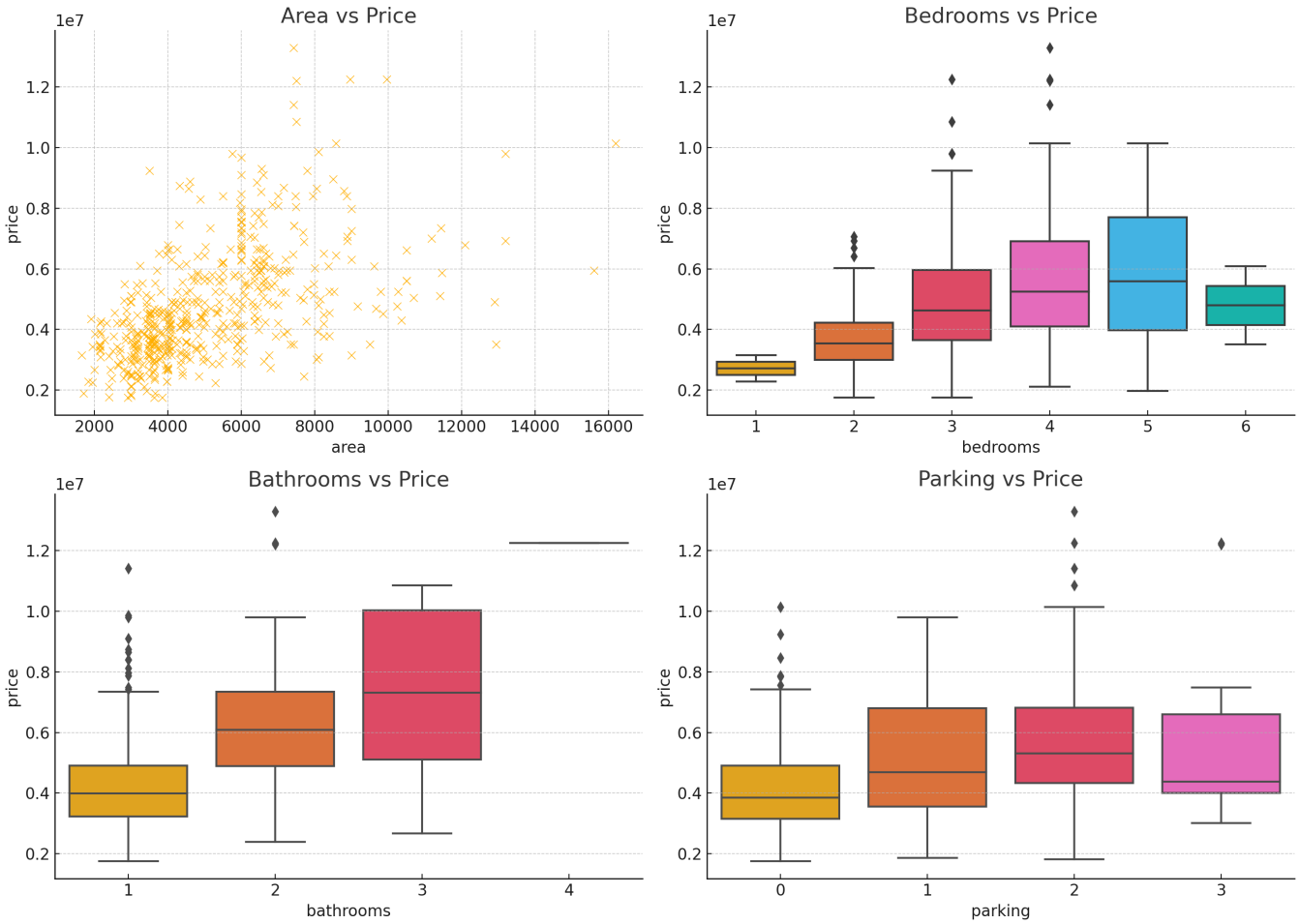
1. Data Loading & Exploration:
   - Read dataset and inspect using head(), info(), describe().
2. Data Cleaning:
   - Handled missing values and converted categories if needed.
3. Visualization:
   - Correlation matrix and scatterplots to understand relationships.
4. Feature Selection:
   - Removed irrelevant or highly correlated features.
5. Model Building:
   - Split data, trained model using LinearRegression.
6. Model Evaluation:
   - Analyzed R^2, Adjusted R^2, residuals, and VIF for multicollinearity.

Data Visualization

Housing Case Study: Multiple Linear Regression



Housing Case Study: Multiple Linear Regression



Results

A robust regression model was developed that included statistically significant variables. The coefficients provided insight into how each factor impacts the property price. The model demonstrated a strong fit, indicated by a high R^2 score.

Key Learnings

- Multiple linear regression can effectively model and predict property prices.
- Ensuring assumptions such as linearity and no multicollinearity is crucial.
- Thoughtful feature selection improves both model performance and interpretability.