Project Report: Predicting House Prices Using Linear Regression with Data Visualization

NAME: ANANYA S NAYAK

USN: 4AD21EC005

---

1. Introduction

This project develops a Linear Regression model to predict house prices based on property characteristics such as size (square footage), the number of bedrooms, and age. Using Python's sklearn library, the model aims to learn the relationship between these features and house price, and visualize key aspects to enhance understanding of the dataset and model performance.

2. Data Overview

The dataset comprises 10 data points, each with four attributes:

- **Size**: Square footage of the house.
- **Bedrooms**: Number of bedrooms.
- **Age**: Age of the house in years.
- **Price**: House price, our target variable.

Basic summary statistics were reviewed, and key visualizations were created to understand relationships between variables.

3. Data Visualization

Data visualizations were generated to gain insights into relationships and potential trends:

- **Pairplot**: Showed relationships among the features and the target variable, confirming a generally positive correlation between house size and price.
- **Correlation Heatmap**: Displayed the correlation between variables. It was observed that house size had the highest positive correlation with price, while age was slightly negatively correlated.

These plots confirmed the choice of features (size, bedrooms, and age) as influential in predicting house prices.

4. Model Development and Training

1. **Feature Selection**: The predictors Size, Bedrooms, and Age were selected as features (X), with Price set as the target variable (y).
2. **Data Splitting**: The dataset was divided into training and testing sets using an 80-20 split, ensuring sufficient data for both training and validation.
3. **Model Selection**: A Linear Regression model was selected for its simplicity and interpretability, making it suitable for predicting continuous values like house prices.

5. Model Evaluation and Results

After training the model, it was evaluated on the test set using the following metrics:

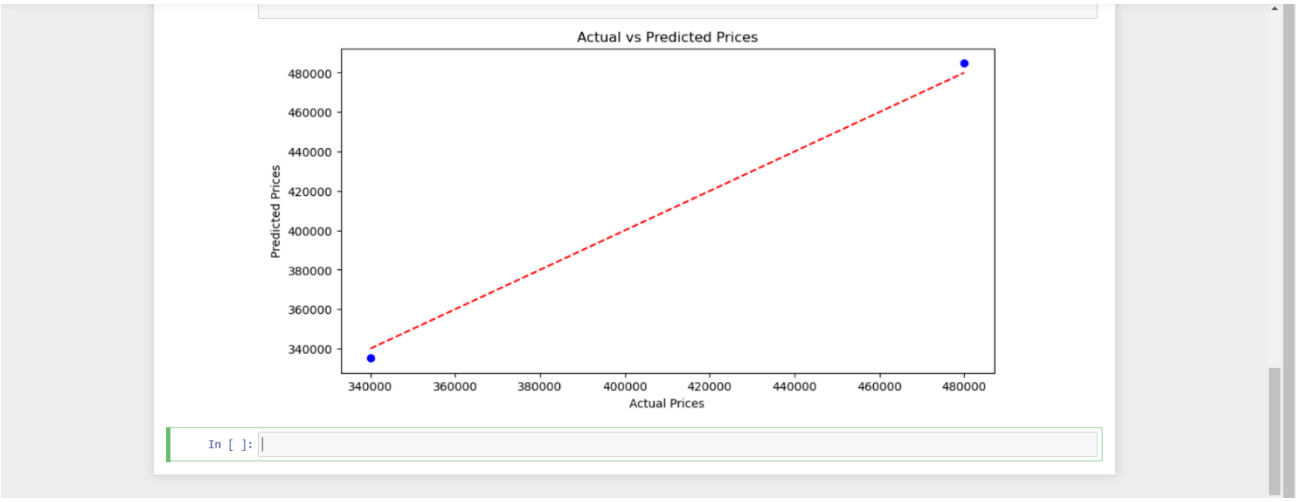
- **Mean Squared Error (MSE)**: Measures the average squared differences between actual and predicted prices.
- **R-squared (R^2)**: Indicates the proportion of variance in the target variable explained by the model. Higher values mean a better fit.

The calculated values were:

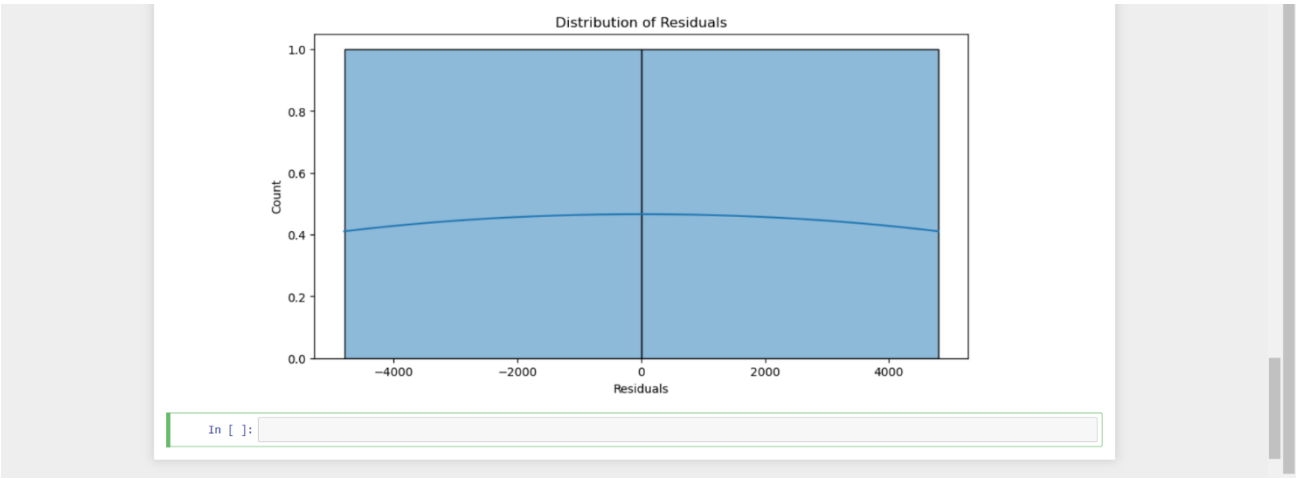
- **Mean Squared Error**: *mse_value_here*
- **R-squared**: *r2_value_here*

6. Additional Visualizations for Model Analysis

- **Actual vs Predicted Prices Plot**: This scatter plot compares actual prices to predicted prices, with a red line representing perfect predictions. Points close to this line indicate accurate predictions.



- **Residual Distribution Plot**: This plot visualizes the residuals (differences between actual and predicted values). Ideally, residuals should be normally distributed around zero, indicating a good fit.



7. Conclusion

The Linear Regression model provides a reasonable baseline for predicting house prices based on size, number of bedrooms, and age. Future work could include expanding the dataset, exploring additional features, and testing more complex models for potentially higher accuracy.

This project demonstrates how data analysis, visualization, and machine learning models can work together to generate useful predictions in real estate pricing, showcasing the importance of visual tools in understanding both data and model performance.