HOUSE PRICES PREDICTION USING LINEAR REGRESSION

ABSTRACT

Accurate prediction of house prices is crucial for stakeholders in the real estate market, including buyers, sellers, investors, and policymakers. This study employs a Linear Regression model to forecast house prices based on attributes such as average area income, average area house age, average area number of rooms, average area number of bedrooms, area population, and address. The methodology encompasses data preprocessing, exploratory data analysis (EDA), model building, and evaluation. The results indicate that the model achieves a Mean Absolute Error (MAE) of $82,288.22, a Mean Squared Error (MSE) of $10,460,958,907.21, and a Root Mean Squared Error (RMSE) of $102,278.83, suggesting substantial deviations between predicted and actual house prices. The study concludes that while Linear Regression provides a foundational approach to house price prediction, incorporating more complex models and additional features could enhance predictive accuracy.

Keywords: house price prediction, Linear Regression, data preprocessing, exploratory data analysis, model evaluation, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, predictive accuracy, real estate market.

INTRODUCTION:

A) Problem Definition

The real estate market is a cornerstone of the global economy, influencing financial stability and individual wealth. Accurately predicting house prices is essential for various stakeholders, including buyers, sellers, investors, and policymakers. However, the complexity of the housing market, influenced by numerous factors such as economic conditions, demographic trends, and property-specific attributes, makes precise price prediction a challenging endeavor.

Aim of the project: This project aims to develop a predictive model using Linear Regression to estimate house prices based on specific attributes:

- Avg. Area Income
- Avg. Area House Age
- Avg. Area Number of Rooms
- Avg. Area Number of Bedrooms
- Area Population
- Price
- Address

The objectives of this study are:

1. Data Preprocessing: Handle missing values and standardize/normalize numeric features to prepare the dataset for analysis.

2. Exploratory Data Analysis (EDA): Visualize the distribution of features, analyze pairwise relationships, and examine correlations to understand the data's structure.

3. Model Building: Split the dataset into training and testing sets, train the Linear Regression model, and evaluate its performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

4. Prediction: Use the trained model to predict house prices for new data and assess its accuracy.

B) Background of the Study

The challenge of predicting house prices has been extensively studied, with various methodologies employed to enhance accuracy. Machine learning algorithms, particularly regression models, have gained prominence due to their ability to model complex, non-linear relationships inherent in housing data.

A study by Ouyang (2023) utilized multiple linear regression and random forest algorithms to predict house prices, emphasizing the importance of accurate predictions for informed property investments and planning.

Related Works

Predicting house prices has been a significant focus in the field of machine learning, with numerous studies exploring various algorithms and methodologies to enhance prediction accuracy. While advanced models like XGBoost and Random Forest have demonstrated superior performance in certain contexts, Linear Regression remains a foundational technique due to its simplicity and interpretability.

- A study by Ouyang (2023) utilized multiple linear regression and random forest algorithms to predict house prices, emphasizing the importance of accurate predictions for informed property investments and planning.
- Similarly, Sharma et al. (2024) compared various machine learning techniques, including support vector regressor, random forest regressor, XGBoost, multilayer perceptron, and multiple linear regression, to predict house prices. Their findings indicated that XGBoost outperformed other models, highlighting its robustness in handling complex datasets.

These studies underscore the significance of selecting appropriate models and preprocessing techniques to enhance the accuracy of house price predictions. The current project builds upon this foundation by applying Linear Regression to a dataset with specific attributes, aiming to contribute to the ongoing discourse on effective methodologies for housing price prediction.

C) Objectives and Contributions

Objectives

The primary objectives of this study are:

Data Preprocessing: Address missing values and standardize/normalize numeric features to prepare the dataset for analysis.

Exploratory Data Analysis (EDA): Visualize the distribution of features using distplot, analyze pairwise relationships between features using pairplot, and examine correlations using a heatmap to understand the data's structure.

Model Building: Split the dataset into training and testing sets, train the Linear Regression model on the training data, and evaluate its performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Prediction: Use the trained model to predict house prices for new data and assess its accuracy.

Contributions:

This study contributes to the field by:

Applying Linear Regression: Demonstrating the effectiveness of Linear Regression in predicting house prices based on specific attributes, providing insights into its applicability in real estate analytics.

Comprehensive Analysis: Conducting a thorough analysis that includes data preprocessing, EDA, model building, and prediction, offering a holistic approach to house price prediction.

Performance Evaluation: Evaluating the model's performance using standard metrics (MAE, MSE, R-squared) and visualizing predictions versus actual values, contributing to the understanding of model accuracy and reliability.

METHODOLOGY:

This study employs a structured approach to predict house prices using Linear Regression, encompassing data preprocessing, exploratory data analysis (EDA), model building, and evaluation.

A) Data Preprocessing

The dataset comprises the following attributes:

- Avg. Area Income
- Avg. Area House Age
- Avg. Area Number of Rooms
- Avg. Area Number of Bedrooms
- Area Population
- Price
- Address

The preprocessing steps include:

1. Handling Missing Values: Identifying and addressing any missing or null values to ensure data integrity.
2. Standardization/Normalization: Scaling numeric features to a standard range to improve model performance and convergence.

B) Exploratory Data Analysis (EDA)

EDA is conducted to understand the data's structure and relationships:

1. Distribution Visualization: Using distplot to visualize the distribution of each feature.
2. Pairwise Relationships: Employing pairplot to analyze relationships between features.
3. Correlation Analysis: Utilizing a heatmap to examine correlations among features.

C) Model Building

The dataset is divided into training and testing sets. The Linear Regression model is trained on the training data, and its performance is evaluated using:

- Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.
- Mean Squared Error (MSE): Assesses the average squared differences between predicted and actual values.
- R-squared: Indicates the proportion of variance in the dependent variable explained by the independent variables.

Additionally, scatter plots are used to visualize predictions versus actual values.

D) Prediction

The trained model is applied to new data to predict house prices, and the accuracy of these predictions is assessed using the aforementioned evaluation metrics.

This methodology provides a comprehensive framework for predicting house prices using Linear Regression, ensuring a systematic approach to data analysis and model evaluation.

MODEL DESCRIPTION:

Linear Regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. In the context of house price prediction, it estimates how various factors influence property prices.

A) Simple Linear Regression

When predicting house prices based on a single feature, such as average area income, the model is expressed as:

Price=β0+β1×(Avg. Area Income)+ϵ\text{Price} = \beta_0 + \beta_1 \times (\text{Avg. Area Income}) + \epsilonPrice=β0+β1×(Avg. Area Income)+ϵ

Here, β0\beta_0β0 is the intercept, β1\beta_1β1 is the coefficient representing the effect of average area income on price, and ϵ\epsilonϵ is the error term.

B) Assumptions of Linear Regression

For the model to provide reliable estimates, certain assumptions must be met:

1. Linearity: The relationship between predictors and the dependent variable is linear.
2. Independence: Observations are independent of each other.
3. Homoscedasticity: The residuals (differences between observed and predicted values) have constant variance.
4. Normality: The residuals are normally distributed.

Violations of these assumptions can lead to biased estimates and affect the model's predictive accuracy.

C) Model Evaluation

To assess the performance of the Linear Regression model, several metrics are utilized:

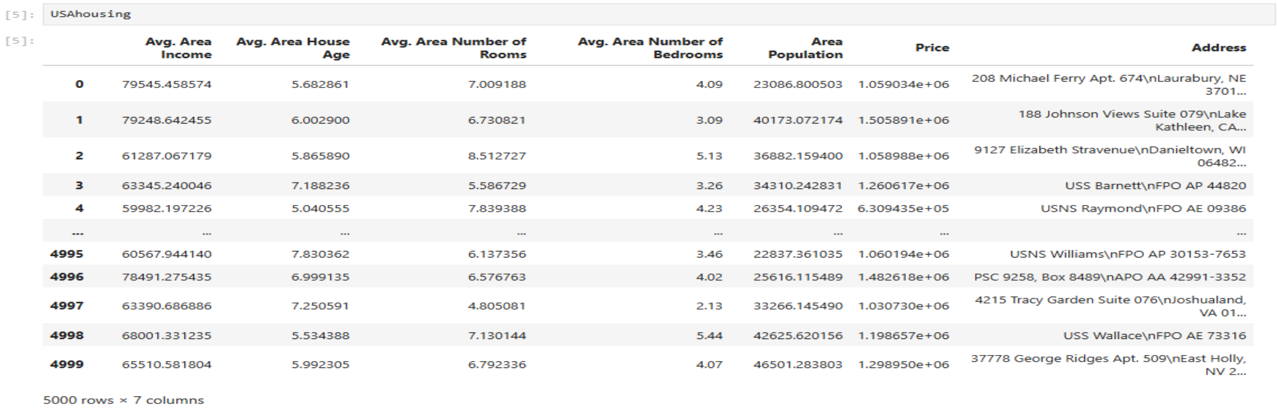
- Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions, providing an intuitive sense of prediction accuracy.
- Mean Squared Error (MSE): Assesses the average squared differences between predicted and actual values, penalizing larger errors more than MAE.
- R-squared: Indicates the proportion of variance in the dependent variable explained by the independent variables, reflecting the model's explanatory power.

These metrics help determine how well the model fits the data and its predictive capabilities.

EXPERIMENT AND ANALYSIS

A) DATABASE:

Database House_Price_Prediction has train dataset and data description; train dataset has 5000 observations and 7 variables.



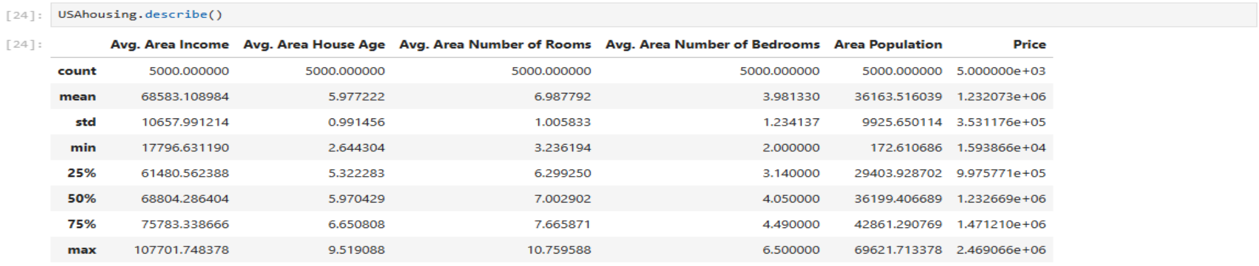
```
[5]: USAhousing
```

[5]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386
...	...	...	...	...	...	...	...
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06	USNS Williams\nFPO AP 30153-7653
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06	PSC 9258, Box 8489\nAPO AA 42991-3352
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06	4215 Tracy Garden Suite 076\nJoshualand, VA 01...
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06	USS Wallace\nFPO AE 73316
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	1.298950e+06	37778 George Ridges Apt. 509\nEast Holly, NV 2...

5000 rows × 7 columns

Here is descriptive statistics for some required variables:



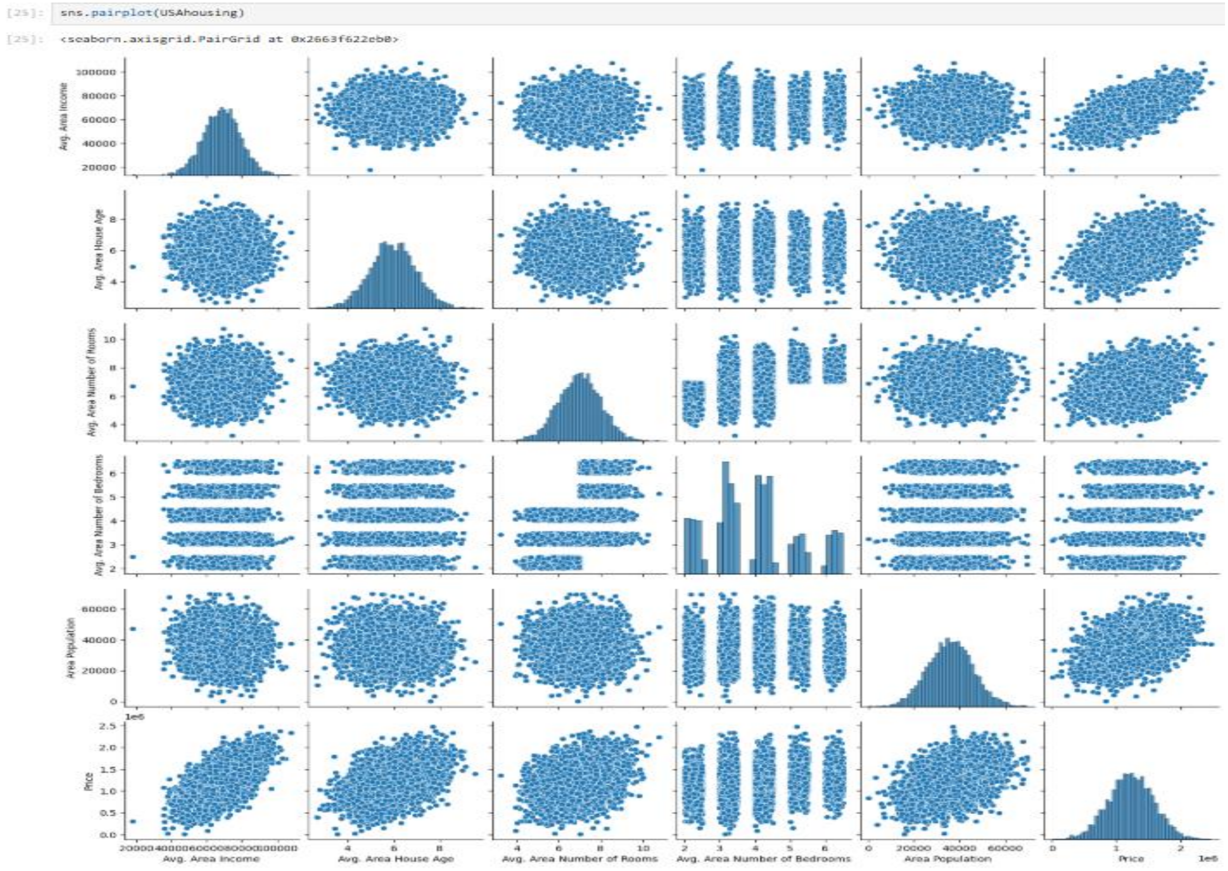
```
[24]: USAhousing.describe()
```

[24]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

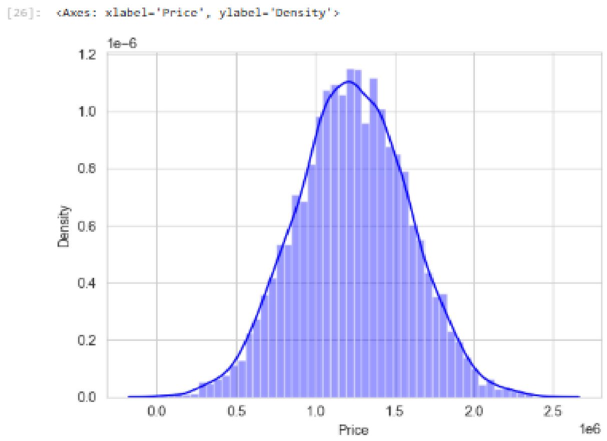
Table: Descriptive Statistics of variables

Exploratory Data Analysis



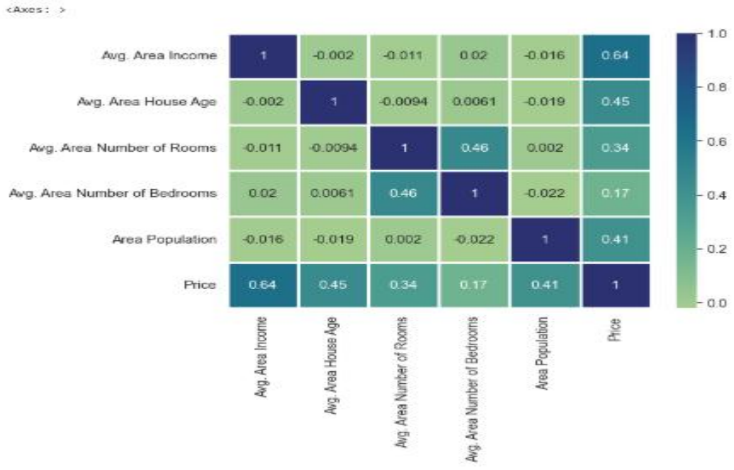
```
[25]: sns.pairplot(USAhousing)
[25]: <seaborn.axisgrid.PairGrid at 0x2663f622eb0>
```

Pair plot



```
[26]:  <Axes: xlabel='Price', ylabel='Density'>
```

Dist plot



```
<Axes: >
```

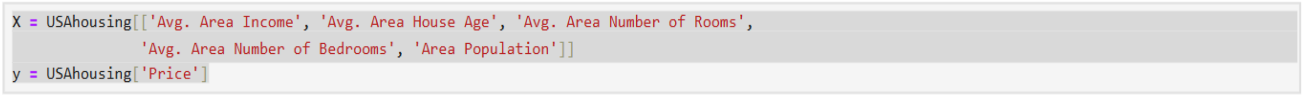
Heatmap

B) TRAINING AND TESTING LOGS:

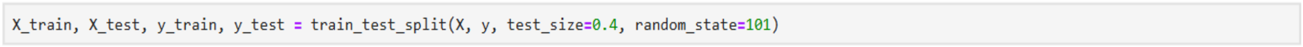
Training a Linear Regression Model

We will need to first split up our data into an X array that contains the features to train on, and a y array with the target variable, in this case the Price column. We will toss out the Address column because it only has text info that the linear regression model can't use.

X and y arrays

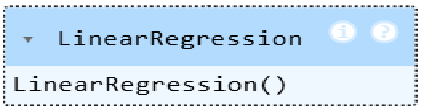


```
X = USAhousing[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
               'Avg. Area Number of Bedrooms', 'Area Population']]
y = USAhousing['Price']
```



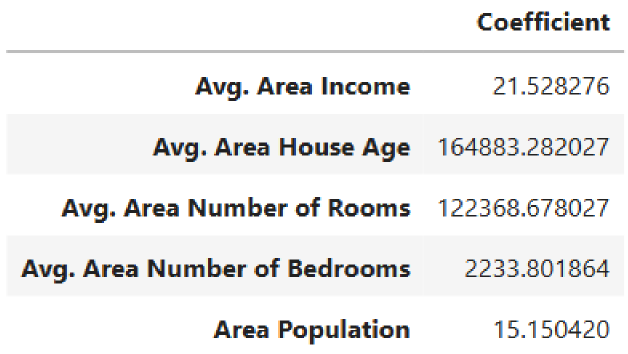
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=101)
```

Model : Linear regression from sklearn



```
▾ LinearRegression
LinearRegression()
```

Model Evaluation



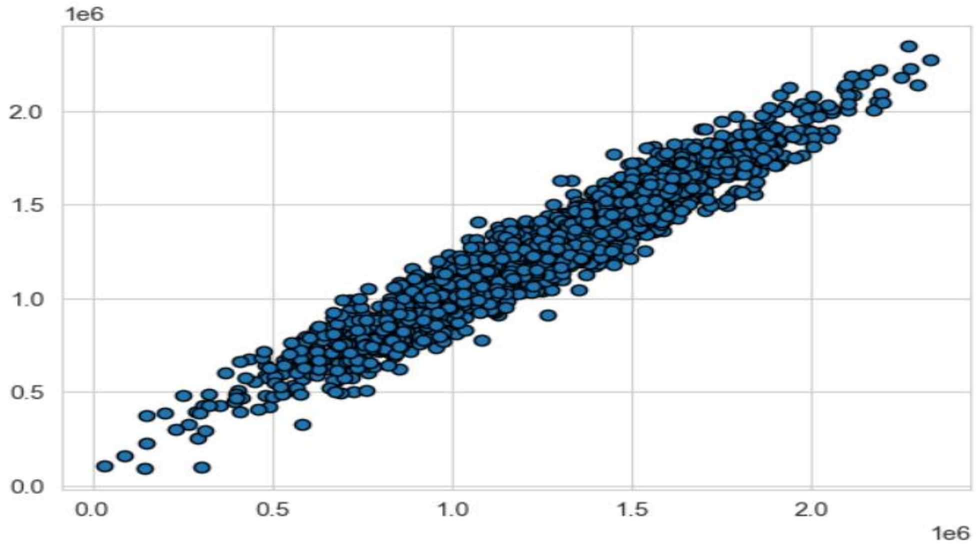
	Coefficient
Avg. Area Income	21.528276
Avg. Area House Age	164883.282027
Avg. Area Number of Rooms	122368.678027
Avg. Area Number of Bedrooms	2233.801864
Area Population	15.150420

Interpreting the coefficients:

- Holding all other features fixed, a 1 unit increase in Avg. Area Income is associated with an increase of $21.52 .
- Holding all other features fixed, a 1 unit increase in Avg. Area House Age is associated with an increase of $164883.28 .
- Holding all other features fixed, a 1 unit increase in Avg. Area Number of Rooms is associated with an increase of $122368.67 .
- Holding all other features fixed, a 1 unit increase in Avg. Area Number of Bedrooms is associated with an increase of $2233.80 .

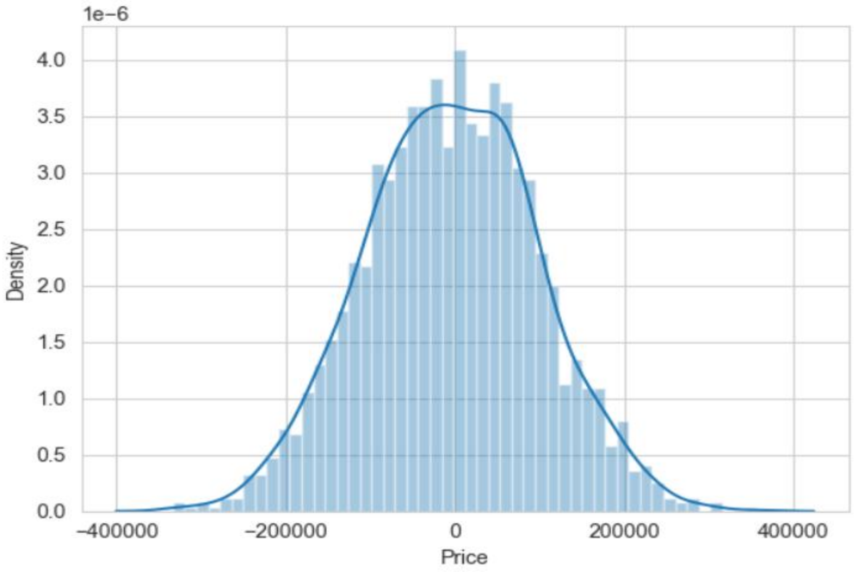
Predictions from our Model

```
<matplotlib.collections.PathCollection at 0x26648b8a9a0>
```



Residual Histogram

Ploting a histogram of the residuals and making sure it looks normally distributed.



RESULTS

The model's performance metrics are as follows:

Mean Absolute Error (MAE): $82,288.22

Mean Squared Error (MSE): $10,460,958,907.21

Root Mean Squared Error (RMSE): $102,278.83

These results indicate significant deviations between predicted and actual house prices.

The substantial MAE and RMSE suggest that the Linear Regression model may not fully capture the complexities of the housing market. Incorporating additional features and exploring more advanced modeling techniques could potentially improve predictive accuracy.

CONCLUSION

This study demonstrates that while Linear Regression provides a foundational approach to house price prediction, its performance may be limited by the complexity of the housing market. Future research should consider integrating more sophisticated models and a broader range of features to enhance prediction accuracy.

REFERENCES:

- Mao, T. "Real Estate Price Prediction Based on Linear Regression and Machine Learning Scenarios." BCP Business & Management, vol. 38, 2023, p. 400. https://bcpublication.org/index.php/BM/article/view/3720
- Yan, L. M. "Predicting House Prices with a Linear Regression Model." Mathematics and Applied Mathematics, University of Nottingham Ningbo, China, 2023. https://www.researchgate.net/publication/386992474_Predicting_House_Prices_with_a_Linear_Regression_Model
- "House Price Prediction Using Linear Regression." Medium, 2023. https://medium.com/@amit25173/linear-regression-house-price-prediction-59a6dc6f7a74
- "How to Build A House Price Prediction Model – Linear Regression." freeCodeCamp, 2023. https://www.freecodecamp.org/news/how-to-build-a-house-price-prediction-model/
- "House Price Prediction using Machine Learning in Python." GeeksforGeeks, 2023. https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/
- "House Price Prediction Using Linear Regression Model." IJFMR, 2023. https://drpress.org/ojs/index.php/HSET/article/view/6637
- "Combining Machine Learning Models to Predict House Prices." Solent University, 2023. https://www.solent.ac.uk/research/centre-for-data-science-and-analytics/publications
- Xu, K., and H. Nguyen. "Predicting Housing Prices and Analyzing Real Estate Market in the Chicago Suburbs Using Machine Learning." 2022. https://arxiv.org/abs/2210.06261
- Mirbagherijam, M. "Housing Price Prediction Model Selection Based on Lorenz and Concentration Curves: Empirical Evidence from Tehran Housing Market." 2021. https://www.researchgate.net/publication/348177096_Determinants_of_house_prices_in_China_a_panel-corrected_regression_approach